Conference Programme and Book of Abstracts
17th Annual Conference of the European Association for Machine Translation

EAMT2014

Dubrovnik, Croatia, 16th-18th June 2014

EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

Edited by Marko Tadić, Philipp Koehn, Johann Roturier, Andy Way

17th Annual Conference of the
European Association for Machine Translation

EAMT 2014

Conference Programme
Book of Abstracts

Edited by
Marko Tadić, Philipp Koehn,
Johann Roturier, Andy Way

Croatian Language
Technologies Society

June 15th–18th 2014
Centre for Advanced Academic Studies
Dubrovnik, Croatia

EAMT2014 Committees

Research Committee

Eleftherios Avramidis (DFKI, Germany)
Alexandra Birch (University of Edinburgh, Scotland)
Ondřej Bojar (Charles University, Czech Republic)
Christian Buck (University of Edinburgh, Scotland)
Bill Byrne (University of Cambridge, England)
Michael Carl (Copenhagen Business School, Denmark)
Francisco Casacuberta (Polytechnic University of Valencia, Spain)
Mauro Cettolo (FBK, Italy)
David Chiang (University of Southern California, USA)
Nadir Durrani (University of Edinburgh, Scotland)
Chris Dyer (Carnegie Mellon University, USA)
Christian Federmann (Microsoft, USA)
Marcello Federico (FBK, Italy)
Mark Fishel (University of Zurich, Switzerland)
Mikel Forcada (Universitat d'Alacant, Spain)
George Foster (NRC, Canada)
Josef van Genabith (CNGL, Dublin City University, Ireland)
Ulrich Germann (University of Edinburgh, Scotland)
Barry Haddow (University of Edinburgh, Scotland)
Rejwanul Haque (Lingo24, UK)
Christian Hardmeier (University of Uppsala, Sweden)
Teresa Herrmann (Karlsruhe Institute of Technology, Germany)
Matthias Huck (University of Edinburgh, Scotland)
Gonzalo Iglesias (University of Cambridge, England)
Jie Jiang (Applied Language Solutions, UK)
Marcin Junczys-Dowmunt (Adam Mickiewicz University, Poland)
Maxim Khalilov (TAUS)
Roland Kuhn (NRC, Canada)
Qun Liu (CNGL, Dublin City University, Ireland)
Adam Lopez (Johns Hopkins University, USA)
José B. Mariño (Polytechnic University of Catalonia, Spain)
Christof Monz (University of Amsterdam, Netherlands)
Jan Niehues (Karlsruhe Institute of Technology, Germany)
Sergio Penkale (Lingo24, UK)
Maja Popović (DFKI, Germany)

Stefan Riezler (University of Heidelberg, Germany)
Felipe Sánchez Martínez (Universitat d'Alacant, Spain)
Khalil Simaan (University of Amsterdam, Netherlands)
Ankit Srivastava (CNGL, Dublin City University, Ireland)
Sara Stymne (University of Uppsala, Sweden)
Jörg Tiedemann (University of Uppsala, Sweden)
Dan Tufiş (Romanian Academy, Romania)
Francis M. Tyers (Universitat d'Alacant, Spain)
David Vilar (Pixformance)
Jörn Wübker (RWTH Aachen, Germany)
François Yvon (LIMSI, France)

User Committee

Jeff Allen (SAP, France)
Nora Aranberri (University of the Basque Country, Spain)
Pratyush Banerjee (Symantec, Ireland)
Nuria Bel (UPF, Spain)
Ergun Bicici (CNGL, Dublin City University, Ireland)
Frédéric Blain (Université du Maine, France)
Bianka Buschbeck (Systran, France)
Tony Clarke (CLS Communication, Switzerland)
Béatrice Daille (University of Nantes, France)
Heidi Depraetere (Cross Language, Belgium)
Ilse Depraetere (Université de Lille III, France)
Mike Dillinger (Translation Optimization Partners, US)
Stephen Doherty (CNGL, Ireland)
Marc Dymetman (Xerox, France)
Andreas Eisele (EC, Luxemboug)
Ray Flournoy (Adobe Systems, US)
Jesús Giménez (Nuance, Spain)
Steve Götz (CNGL, Ireland)
Daniel Grasmick (Lucy Software and Services, Germany)
Declan Groves (Microsoft, Ireland)
Ana Guerberof (Universitat Autònoma de Barcelona, Spain)
Rafael Guzman (Symantec, Ireland)
Olivier Hamon (Syllabs, France)
Viggo Hansen (EAMT Executive Committee, Denmark)
Fred Hollowood (Fred Hollowood Consulting, Ireland)

Dorothy Kenny (CNGL, Dublin City University, Ireland)
Qun Liu (CNGL, Dublin City University, Ireland)
John Moran (CNGL, Ireland)
Sharon O'Brien (CNGL, Dublin City University, Ireland)
Sergio Pelino (Google, US)
Sergio Penkale (Lingo24, UK)
Mirko Plitt (Modulo Language Automation, Switzerland)
Bruno Pouliquen (WIPO, Switzerland)
Alexandros Poulis (Lionbridge, Finland)
Manny Rayner (University of Geneva, Switzerland)
Rubén Rodríguez de la Fuente (Paypal)
Raphael Rubino (Prompsit Language Engineering, Spain)
Marta Ruiz Costa-Jussà (Catalan Politechnic University, Spain)
Dag Schmidtke (Microsoft Ireland, Dublin, Ireland)
Jean Senellart (Systran, France)
Violeta Seretan (University of Geneva, Switzerland)
Svetlana Sheremetyeva (LanA Consulting, ApS, Denmark; South Ural State University, Russia)
Yanli Sun (Symantec, China)
Midori Tatsumi (Translator, Japan)
Chris Wendt (Microsoft Research, US)
Francois Yvon (University Paris Sud 11, France)
Ventsislav Zhechev (Autodesk, Switzerland)
Jost Zetsche (International Writers' Group, US)
Andy Way (CNGL, Dublin City University, Ireland)

Organizing Committee

Božo Bekavac (University of Zagreb, Croatia)
Philipp Koehn (University of Edinburgh, UK), Research Track Program Chair
Johann Roturier (Symantec, Ireland), User Track Program Chair
Marko Tadić (University of Zagreb, Croatia), Chair
Andy Way (CNGL, Dublin City University, Ireland), Project/Product Track Program Chair

Acknowledgments

The European Association for Machine Translation acknowledges with gratitude the support and sponsoring of the following institutions and companies:

Bloomberg

Bloomberg (Silver sponsor)

European Language Resources Agency ELRA (Bronze sponsor)

Springer

Springer (Best Paper Award sponsor)

ULIX

Ulix Travel Agency

Ministry of Science, Education and Sport of the Republic of Croatia

Hrvatsko društvo za jezične tehnologije / Croatian Language Technologies Society

EAMT2014 Programme at a Glance

EAMT2014 program	Sunday, 2014-06-15	Monday, 2014-06-16	Tuesday, 2014-06-17	Wednesday, 2014-06-18
08:30-09:00		Registration		
09:00-09:30		Welcome Remarks	Slot 6: Scarton & Specia	Slot 11: Castilho et al.
09:30-10:00		Invited talk	Slot 7: Shah & Specia	Slot 12: Popović et al.
10:00-10:30			Slot 8: Ngoc et al.	Slot 13: Čulo
10:30-11:00		Poster Boaster A	Poster Boaster B	Slot 14: Skadiņš et al.
11:00-11:30		Coffee break	Coffee break	Coffee break
11:30-12:00				Slot 15: Junczys-Dowmunt & Pouliquen
12:00-12:30	Registration	Poster Session A	Poster Session B	Slot 16: Toral et al.
12:30-13:00				Closing Remarks
13:00-13:30		Lunch (Restaurant Mimoza)	Lunch (Restaurant Mimoza)	
13:30-14:00				
14:00-14:30		Slot 1: El Kholy & Habash	Slot 9: Lommel et al.	
14:30-15:00		Slot 2: Durrani & Koehn	Slot 10: Ruiz & Federico	
15:00-15:30				
15:30-16:00		Coffee break	Best Thesis Award	
16:00-16:30	QTLaunchPad Workshop	Slot 3: Garcia et al.	EAMT General Assembly	
16:30-17:00		Slot 4: Schaefer et al.		
17:00-17:30		Slot 5: Marg & Casanellas		
17:30-18:00				
18:00-18:30				
18:30-19:00		Guided city tour (in front of CAAS)	Excursion + Conference dinner (Ston)	
19:00-19:30				
19:30-20:00				
20:00-20:30	Welcome Reception (CAAS court)			
20:30-21:00				
21:00-21:30				
21:30-22:00				
22:00-22:30				
22:30-23:00				

research track
user track
mixed track

EAMT2014 Detailed Programme

Sunday, 2014-06-15

12:00–14:30 **Registration**

Preconference Workshop: QTLaunchPad

14:30 Opening of the Workshop
14:30–15:00 Bogdan Babych (University of Leeds): Machine translation evaluation for MT development: improving MT quality with evaluation-oriented methods
15:00–15:30 Christian Federmann (Microsoft): Quality translation: where are we now, and where are we going?
15:30–16:00 Kim Harris (text&form): Multidimensional Quality Metrics: the art of the science
16:00–16:30 Maja Popović (German Research Centre for Artificial Intelligence): The limits of automation in MT evaluation

16:30–17:00 **Coffee Break**

17:00–17:30 Lucia Specia (University of Sheffield): Predicting human and machine translation quality
17:30–18:00 Antonio Toral and Federico Gaspari (CNGL, Dublin City University): Source-language phenomena triggering MQM errors
18:00–18:30 Jost Zetzsche (International Writers' Group): Are quality metrics for machine translation the way to the heart of the translator?
18:30–19:00 Final discussion on "The limits of automation in MT evaluation and quality estimation"

20:00–22:00 **EAMT2014 Welcome Reception** (CAAS Court)

This half-day Workshop has been generously supported by the European Association for Machine Translation (EAMT).

Further details on the Workshop:
www.qt21.eu/launchpad/content/eamt2014

More information on the European Commission-funded QTLaunchPad project: www.qt21.eu/launchpad/

Monday, 2014-06-16

8:30–9:00 **Registration**

9:00–9:30 **Opening and Welcome Remarks**

9:30–10:30 **Invited talk** (chairperson: Andrew Way)
Jost Zetzsche: Encountering the Unknown (Part 2)

10:30–11:00 **Poster Boaster A** (chairperson: Marko Tadić)

11:00–11:30 **Coffee Break**

11:30–13:00 **Poster Session A** (chairperson: Marko Tadić)

Research track

Rohit Gupta and Constantin Orasan: Incorporating Paraphrasing in Translation Memory Matching and Retrieval

Marion Weller, Alexander Fraser and Ulrich Heid: Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Juan A. Pérez-Ortiz, Felipe Sánchez-Martínez, Mikel L. Forcada and Rafael C. Carrasco: An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words

Germán Sanchis Trilles and Daniel Ortiz-Martínez: Efficient Wordgraph Pruning for Interactive Translation Prediction

Saab Mansour and Hermann Ney: Translation Model Based Weighting for Phrase Extraction

Xingyi Song, Lucia Specia and Trevor Cohn: Data Selection for Discriminative Training in Statistical Machine Translation

Product/Project track

Rafał Jaworski and Renata Ziemlińska: The translaide.pl system: an effective real world installation of translation memory searching and EBMT

Volker Steinbiss: Collaborative Project EU-BRIDGE – Bridges Across the Language Divide

Marcello Federico: MATECAT Project

Aswarth Dara, Josef van Genabith, Qun Liu, John Judge and Antonio Toral: PEDAL: Post-Editing with Dynamic Active Learning

Philipp Koehn, Michael Carl, Francisco Casacuberta and Eva Marcos: CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation

Johann Roturier, Pierrette Bouillon, Laurence Roguet, Barry Haddow and Robert Grabowski: ACCEPT : Automated Community Content Editing PorTal

Olga Beregovaya and David Landan: Source Content Analysis and Training Data Selection Impact on an MT-driven Program Design with a Leading LSP

Raivis Skadiņš, Inguna Skadiņa, Andrejs Vasiļjevs: TAAS: Terminology as a Service

13:00–14:30 **Lunch** (Restaurant Mimoza)

Research track (chairperson: Bogdan Babych)

14:30–15:00 Ahmed El Kholy and Nizar Habash: Alignment Symmetrization Optimization targeting Phrase Pivot Statistical Machine Translation

15:00–15:30 Nadir Durrani and Philipp Koehn: Improving Machine Translation via Triangulation and Transliteration

15:30–16:00 **Coffee Break**

User track (chairperson: Johann Roturier)

16:00–16:30 Mercedes Garcia, Karan Singla, Aniruddha Tammewar, Bartolome Mesa-Lao, Ankita Thakur, Anusuya M. A., Srinivas Banglore and Michael Carl: SEECAT: Speech & Eye-tracking Enabled Computer Assisted Translation

16:30–17:00 Falko Schaefer, Joeri Van de Walle and Joachim Van den Bogaert: Moses SMT as an Aid to Translators in the Production Process

17:00–17:30 Lena Marg and Laura Casanellas: Assumptions, Expectations and Outliers in Post-Editing

18:00–20:00 **Guided City Tour** (starting in front of CAAS)

Tuesday, 2014-06-17

Research track (chairperson: Khalil Simaan)

9:00–9:30 Carolina Scarton and Lucia Specia: Document-level translation quality estimation: exploring discourse and pseudo-references

9:30–10:00 Kashif Shah and Lucia Specia: Quality estimation for translation selection

10:00–10:30 Ngoc Quang Luong, Laurent Besacier and Benjamin Lecouteux: An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation

10:30–11:00 **Poster Boaster B** (chairperson: Mikel Forcada)

11:00–11:30 **Coffee Break**

11:30–13:00 **Poster Session B** (chairperson: Mikel Forcada)

Product/Project track

Orero Pilar, Carla Ortiz-Boix and Anna Matamala: HBB4ALL: media accessibility in HbbTV

Anna Matamala and Carla Ortiz-Boix: Technologies for linguistic and sensorial accessibility: the ALST project

António Branco and Petya Osenova: QTLeap – Quality Translation with Deep Language Engineering Approaches

Fatiha Sadat: The ASMAT project – Arabic Social Media Analysis Tools

Marko Tadić: XLike: Cross-lingual Knowledge Extraction

Antonio Toral, Tommi Pirinen, Andy Way, Raphaël Rubino, Gema Ramírez-Sánchez, Sergio Ortiz, Miquel Esplà, Felipe Sánchez-Martínez, Mikel Forcada, Nikola Ljubešić, Prokopis Prokopidis and Vasilis Papavasiliou: Abu-MaTran: Automatic building of Machine Translation

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris and Hans Uszkoreit: QTLaunchPad

Vincent Vandeghinste and Ineke Schuurman: Able-to-Include: Improving Accessibility for people with Intellectual Disabilities

Vincent Vandeghinste, Tom Vanallemeersch, Véronique Hoste, Marie-Francine Moens, Patrick Wambacq, Karin Coninx and Ken De Wachter: Smart Computer Aided Translation Environment

User track

Sharon O'Brien, Joss Moorkens and Joris Vreeke: Kanjingo – A Mobile App for Post-Editing

Federico Fancellu and Andy Way: Standard language variety conversion using SMT

Vicent Alabau and Luis A. Leiva: Collaborative Web UI Localization, or How to Build Feature-rich Multilingual Datasets

Johann Roturier, David Silva and Linda Mitchell: Using the ACCEPT framework to conduct an online community-based translation evaluation study

Mark Fishel and Rico Sennrich: Handling Technical OOVs in SMT

13:00–14:30 **Lunch** (Restaurant Mimoza)

Research track (chairperson: Lucia Specia)

14:30–15:00 Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis and Hans Uszkoreit: Using a new analytic measure for the annotation and analysis of MT errors on real data

15:00–15:30 Nicholas Ruiz and Marcello Federico: Complexity of Spoken Versus Written Language for Machine Translation

15:30–16:00 **Best Thesis Award**

Gennadi Lembersky: The Effect of Translationese on Statistical Machine Translation

16:00–17:00 **EAMT General Assembly**

17:15–23:00 **Excursion with Conference Dinner**

(starting in front of Hilton hotel)

Wednesday, 2014-06-18

Research track (chairperson: Philipp Koehn)

9:00–9:30 Sheila Castilho, Sharon O'Brien and Fabio Alves: Does post-editing increase usability? A study with Brazilian Portuguese as Target Language

9:30–10:00 Maja Popović, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis and Hans Uszkoreit: Relations between different types of post-editing operations, cognitive effort and temporal effort

10:00–10:30 Oliver Čulo: Approaching Machine Translation from Translation Studies – a perspective on commonalities, potentials, differences

User track (chairperson: James Hodson)

10:30–11:00 Raivis Skadiņš, Inguna Skadiņa, Mārcis Pinnis, Andrejs Vasiļjevs and Tomáš Hudík: Application of Machine Translation in Localization into low-resourced languages

11:00–11:30 **Coffee Break**

11:30–12:00 Marcin Junczys-Dowmunt and Bruno Pouliquen: SMT of German Patents at ORGNAME: Decompounding and Verb Structure Pre-reordering

12:00–12:30 Antonio Toral, Raphael Rubino, Miquel Esplà, Tommi Pirinen, Andy Way and Gema Ramírez-Sánchez: Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain

12:30–13:00 **Closing Remarks and Conference Closing**

Abstracts

Invited talk
Monday, 2014-06-16, 9:30–10:30

Jost Zetzsche: Encountering the Unknown, Part 2

At the AMTA conference four years ago in Denver, I challenged both translators and the MT community by presenting them with "task lists" of items that would help them build bridges to each other.
The tasks that the translators were "charged" with were to look back at previous responses to technology, put into perspective what MT is in relation to other technologies, differentiate between different forms of MT, employ MT where appropriate, and embrace their whole identity.
The MT community was asked to acknowledge the origin of data and linguistic expertise it uses, communicate in terms that are down to earth and truthful, engage the translation community in meaningful ways, listen to the translation community, and embrace their whole identity.
For this presentation I will attempt to evaluate how the two sides have done, what other tasks might need to be added, and whether there actually are still two sides.
I have collected feedback from the greater community of translators for this presentation.

Research track
Poster Session A: Monday, 2014-06-16, 11:30–13:00

Rohit Gupta and Constantin Orăsan: Incorporating Paraphrasing in Translation Memory Matching and Retrieval

Current Translation Memory (TM) systems work at the surface level and lack semantic knowledge while matching. This paper presents an approach to incorporating semantic knowledge in the form of paraphrasing in matching and retrieval. Most of the TMs use Levenshtein edit-distance or some variation of it. Generating additional segments based on the paraphrases available in a segment results in exponential time complexity while matching. The reason is that a particular phrase can be paraphrased in several ways and there can be several possible phrases in a segment which can be paraphrased. We propose an efficient approach to incorporating paraphrasing with edit-distance. The approach is based on greedy approximation and dynamic programming. We have obtained significant improvement in both retrieval and translation of retrieved segments for TM thresholds of 100%, 95% and 90%.

Marion Weller, Alexander Fraser and Ulrich Heid: Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT

Translating in technical domains is a well known problem in SMT, as the lack of parallel documents causes significant problems of sparsity. We discuss and compare different strategies for enriching SMT systems built on general domain data with bilingual terminology mined from comparable corpora. In particular, we focus on the target language inflection of the terminology data and present a pipeline that can generate previously unseen inflected forms.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Juan A. Pérez-Ortiz, Felipe Sánchez-Martínez, Mikel L. Forcada and Rafael C. Carrasco: An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words

A method is presented to assist users with no background in linguistics in adding the unknown words in a text to monolingual dictionaries such as those used in rulebased machine translation systems. Adding a word to these dictionaries requires identifying its stem and the inflection paradigm to be used in order to generate all its word forms. Our method is based on a previous interactive approach in which non-expert users were asked to validate whether some tentative word forms were correct forms of the new word; these validations were then used to determine the most appropriate stem and paradigm. The previous approach was based on a set of intuitive heuristics designed both to obtain an estimate of the eligibility of each candidate stem/paradigm combination and to determine the word form to be validated at each step. Our new approach however uses formal models for both tasks (a hidden Markov model to estimate eligibility and a decision tree to select the word form) and achieves significantly better results.

Germán Sanchis Trilles and Daniel Ortiz-Martínez: Efficient Wordgraph Pruning for Interactive Translation Prediction

When applying interactive translation prediction in real-life scenarios, response time is critical for the users to accept the interactive translation prediction system as a potentially useful tool. In this paper, we report on three different strategies for reducing the computation time required by a state-of-the-art interactive translation prediction system, so that automatic completions are delivered in real time. The best possibility turns out to be to directly prune the wordgraphs derived from the search procedure, achieving real-time response rates without any degradation whatsoever in the quality of the completions provided.

Saab Mansour and Hermann Ney: Translation Model Based Weighting for Phrase Extraction

Domain adaptation for statistical machine translation is the task of altering general models to improve performance on the test domain. In this work, we suggest several novel weighting schemes based on translation models for adapted phrase extraction. To calculate the weights, we first phrase align the general bilingual training data, then, using domain specific translation models, the aligned data is scored and weights are defined over these scores. Experiments are performed on two translation tasks, German-to-English and Arabic-to-English translation with lectures as the target domain. Different weighting schemes based on translation models are compared, and significant improvements over automatic translation quality are reported. In addition, we compare our work to previous methods for adaptation and show significant gains.

Xingyi Song, Lucia Specia and Trevor Cohn: Data Selection for Discriminative Training in Statistical Machine Translation

The efficacy of discriminative training in Statistical Machine Translation is heavily dependent on the quality of the development corpus used, and on its similarity to the test set. This paper introduces a novel development corpus selection algorithm – the LA selection algorithm. It focuses on the selection of development corpora to achieve better translation quality on unseen test data and to make training more stable across different runs, particularly when hand-crafted development sets are not available, and for selection from noisy and potentially non-parallel, large scale web crawled data. LA does not require knowledge of the test set, nor the decoding of the candidate pool before the selection. In our experiments, development corpora selected by LA lead to improvements of over 2.5 BLEU points when compared to random development data selection from the same larger datasets.

Product/Project track
Poster Session A: Monday, 2014-06-16, 11:30–13:00

Rafał Jaworski and Renata Ziemlińska: The translaide.pl system: an effective real world installation of translation memory searching and EBMT

translaide.pl is a CAT system developed by the Polish company PolEng Sp. z o.o. that supports multiple input and output languages. The main idea of the system is to enable the sharing of resources among translators. A demo version of the system is available on the internet (http://translaide.pl), yet it is primarily intended for exclusive use in a single corporation. The system has been successfully implemented in two companies dealing with high-volume content to be translated.

Volker Steinbiss: Collaborative Project EU-BRIDGE – Bridges Across the Language Divide

EU-BRIDGE aims to develop speech and machine translation capabilities that exceed the state-of-the-art in new and more challenging use cases. EU-BRIDGE seeks to achieve rapid technology transition and market insertion by creating a cloud-based speech translation service infrastructure upon which four use cases are built.

Marcello Federico: MATECAT Project

The objective of MateCat is to improve the integration of machine translation and human translation within the so-called computer aided translation (CAT) framework. Several recent studies have shown that post-editing suggestions of a statistical MT engine can substantially improve productivity of professional translators. MateCat leverages the growing interest and expectations in statistical MT by advancing the state of the art along directions that will hopefully accelerate its adoption by the translation industry.

Aswarth Dara, Josef van Genabith, Qun Liu, John Judge and Antonio Toral: PEDAL: Post-Editing with Dynamic Active Learning

Machine translation, in particular statistical machine translation (SMT), is making big inroads into the localisation and translation industry. In typical workflows (S)MT output is checked and manually post-edited by human translators. Recently, a significant amount of research has concentrated on capturing human post-editing outputs as early as possible to incrementally update/modify SMT models to avoid repeat mistakes. Typically in these approaches, MT and post-edits happen sequentially and chronologically, following the way unseen data is presented. In this project, we add to the existing literature addressing the question whether, and if so, to what extent, this process can be improved upon by Active Learning, where input is not presented chronologically but dynamically selected according to criteria that maximise performance with respect to the remaining data. The criteria we use are novel and allow the MT system to improve its performance earlier.

Philipp Koehn, Michael Carl, Francisco Casacuberta and Eva Marcos: CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation

In its third year, the CASMACAT project has developed – in collaboration with the MATECAT project – a new open source workbench for translators that is deployed over the web and as a stand-alone tool. With insights from cognitive studies of translator behavior, new types of assistance have been developed and tested in field trials. The cognitive studies of translator behavior with the CASMACAT workbench include: 1) identification of translator types and styles; 2) cognitive and user models of translation processes. The tool comprising the CASMACAT workbench are currently integrated into a stand-alone version that runs on any computer.

Johann Roturier, Pierrette Bouillon, Laurence Roguet, Barry Haddow and Robert Grabowski: ACCEPT: Automated Community Content Editing PorTal

ACCEPT is a Collaborative Project – STREP aimed at developing new methods and techniques to make machine translation (MT) work better in the environment characterised by internet communities sharing specific information. Today, anyone can in principle create information and make it available to anyone in the world who has Internet access. Yet the language barrier remains: however accessible information is, it is still only available to those who speak the language it is written in. ACCEPT's mission is to help communities share information more effectively across the language barrier by improving the quality of machine-translated community content. The project proposes a new approach to help MT work better for community content, in order to ensure that the result is comprehensible and correct.

Olga Beregovaya and David Landan: Source Content Analysis and Training Data Selection Impact on an MT-driven Program Design with a Leading LSP

Clients requiring translation and localization services have come to require an ever-increasing volume of data to be processed, and an unprecedented diversity in the nature of the data to be translated. To meet the increasing demand for translation and the various requirements to the quality of the target output, nearly all language service providers (LSPs) have added machine translation (MT) and various levels of post editing (PE) as integral components of their service offerings. It has been repeatedly shown that statistical MT engines trained on clean and relevant in-domain data lead to better quality of machine translation output, by using just one of the quality measurement metrics. The importance of corpus preparation and curation and matching the training corpus to the specifics of the content to be translated cannot be overstated. Because of the rapid growth of the amount of data that must be processed, it is imperative that LSPs replace human source content

and training corpora evaluations, which are costly both in terms of time and money spent, with a range of programmatic methodologies, which allow for predicting the quality of machine-translated output when specific training data is used.

Raivis Skadiņš, Inguna Skadiņa, Andrejs Vasiļjevs: TAAS: Terminology as a Service

The project implements a new paradigm in terminology work creating an online platform termunity.com to automate terminology identification, acquisition and processing tasks. The automation of individual tasks is provided as a set of interoperable cloud-based services integrated into workflows. These services automate identification of term candidates in user-provided documents, the lookup of translation equivalents in online terminology resources and on the Web by automatically extracting multilingual terminology from comparable and parallel online resources. Although term identification is very challenging even to human annotators, we can achieve a comparable precision with automatic methods using the extended term tagging system. For example, for Latvian an average precision of 53.8% was achieved in comparison to an average annotator agreement rate of 63.3%. An API is provided for usage of the terminology services and terminology data by external systems. This API-level integration is currently implemented by the memoQ CAT tool and the LetsMT statistical MT system.

Research track
Monday, 2014-06-16, 14:30–15:30

Ahmed El Kholy and Nizar Habash: Alignment Symmetrization Optimization targeting Phrase Pivot Statistical Machine Translation

An important step in mainstream statistical machine translation (SMT) is combining bidirectional alignments into one alignment model. This process is called symmetrization. Most of the symmetrization heuristics and models are focused on direct

translation (source-to-target). In this paper, we present symmetrization heuristic relaxation to improve the quality of phrasepivot SMT (source-[pivot]-target). We show positive results (1.2 BLEU points) on Hebrew-to-Arabic SMT pivoting on English.

Nadir Durrani and Philipp Koehn: Improving Machine Translation via Triangulation and Transliteration

In this paper we improve Urdu→Hindi⇄English machine translation through triangulation and transliteration. First we built an Urdu→ Hindi SMT system by inducing triangulated and transliterated phrase-tables from Urdu–English and Hindi–English phrase translation models. We then use it to translate the Urdu part of the Urdu-English parallel data into Hindi, thus creating an artificial Hindi-English parallel data. Our phrase-translation strategies give an improvement of up to +3.35 BLEU points over a baseline Urdu→ Hindi system. The synthesized data improve Hindi→English system by +0.35 and English→Hindi system by +1.0 BLEU points.

User track
Monday, 2014-06-16, 16:00–17:30

Mercedes Garcia, Karan Singla, Aniruddha Tammewar, Bartolome Mesa-Lao, Ankita Thakur, Anusuya M. A., Srinivas Banglore and Michael Carl: SEECAT: Speech & Eye-tracking Enabled Computer Assisted Translation

Typing has traditionally been the only input method used by human translators working with computer-assisted translation (CAT) tools. However, speech is a natural communication channel for humans and, in principle, it should be faster and easier than typing from a keyboard. This contribution investigates the integration of automatic speech recognition (ASR) in a CAT workbench testing its real use by human translators while post-editing machine translation (MT) outputs. This paper also explores the use of MT

combined with ASR in order to improve recognition accuracy in a workbench integrating eye-tracking functionalities to collect process-oriented information about translators' performance.

Falko Schaefer, Joeri Van de Walle and Joachim Van den Bogaert: Moses SMT as an Aid to Translators in the Production Process

SAP has been heavily involved in the implementation and deployment of machine translation (MT) within the company since the early 1990s. In 2013, SAP initiated an extensive proof of concept project, based on the statistical MT system Moses (Koehn, et al., 2007), in collaboration with the external implementation partner CrossLang. The project focused on the use of Moses SMT as an aid to translators in the production process. This paper describes the outcome of the productivity evaluation for technical documents pertaining to SAP's Rapid Deployment Solutions (RDS), which was performed as part of this proof of concept project.

Lena Marg and Laura Casanellas: Assumptions, Expectations and Outliers in Post-Editing

As a multilingual vendor, we have access to machine translation (MT) scoring and other evaluation data on a wide range of language combinations and content types; we also have experience with different MT systems in production. Our daily work involves the collaboration with a wide spectrum of translation partners, from very MT-savvy to novices. Being exposed to MT in such a varied and large-scale setup, we would like to share some of our insights into assumptions, expectations and outliers observed with regard to MT quality, productivity and suitability with a particular focus on the challenges that post-editor behavior presents in this context. Our observations are based on data correlations carried out at the end of 2013 from a database that contains all evaluation data produced during this year and recent surveys with some of our very MT-savvy translation partners for deeper, locale-specific insights.

Research track
Tuesday, 2014-06-17, 9:00–10:30

Carolina Scarton and Lucia Specia: Document-level translation quality estimation: exploring discourse and pseudo-references

Predicting the quality of machine translations is a challenging topic. Quality estimation (QE) of translations is based on features of the source and target texts (without the need for human references), and on supervised machine learning methods to build prediction models. Engineering well-performing features is therefore crucial in QE modelling. Several features have been used so far, but they tend to explore very short contexts within sentence boundaries. In addition, most work has targeted sentence-level quality prediction. In this paper, we focus on document level QE using novel discursive features, as well as exploiting pseudo-reference translations. Experiments with features extracted from pseudo-references led to the best results, but the discursive features also proved promising.

Kashif Shah and Lucia Specia: Quality estimation for translation selection

We describe experiments on quality estimation to select the best translation among multiple options for a given source sentence. We consider a realistic and challenging setting where the translation systems used are unknown, and no relative quality assessments are available for the training of prediction models. Our findings indicate that prediction errors are higher in this blind setting. However, these errors do not have a negative impact in performance when the predictions are used to select the best translation, compared to non-blind settings. This holds even when test conditions (text domains, MT systems) are different from model building conditions. In addition, we experiment with quality prediction for translations produced by both translation systems and human translators. Although the latter are on average of much higher quality, we show that automatically distinguishing the two types of translation is not a trivial problem.

Ngoc Quang Luong, Laurent Besacier and Benjamin Lecouteux: An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation

During decoding, the SMT decoder travels over all complete paths on the Search Graph (SG), seeks those with cheapest costs and backtracks to read off the best translations. Although these winners beat the rest in model scores, there is no certain guarantee that they have the highest quality with respect to the human references. This paper exploits Word Confidence Estimation (WCE) scores in the second pass of decoding to enhance the MT quality. By using the confidence score of each word in the *N*-best list to update the cost of SG hypotheses containing it, we hope to "reinforce" or "weaken" them relied on word quality. After the update, new best translations are re-determined using updated costs. In the experiments on our *real WCE scores and ideal (oracle) ones*, the latter significantly boosts one-pass decoder by 7.87 BLEU points, while the former yields an improvement of 1.49 points for the same metric.

Product/Project track
Poster Session B: Tuesday, 2014-06-17, 11:30–13:00

Orero Pilar, Carla Ortiz-Boix and Anna Matamala: HBB4ALL: media accessibility in HbbTV

HBB4ALL builds on HbbTV, as the major European standard, for converged services and looks at both the production and service sides. HbbTV 1.x devices are widely available in the market while HbbTV version 2.0 is currently under development. HbbTV provides a straight-forward specification on how to combine broadcast and broadband content plus interactive applications. TV content can be enhanced with additional synchronised services in a personalised manner. For access services this opens an entirely new opportunity for users who may choose an access service delivered via their IP connection which then seamlessly integrates with the regular broadcast programme. The project will test access services

in various pilot implementations and gather user feedback to assess the acceptance and the achievable quality of service in the various delivery scenarios (broadcasting, hybrid, full IP): Multi-platform subtitle services, alternative audio production and distribution, automatic user interface adaptation, and sign-language translation service.

Anna Matamala and Carla Ortiz-Boix: Technologies for linguistic and sensorial accessibility: the ALST project

ALST aims to implement three existing technologies (speech recognition, machine translation and speech synthesis) into two different audiovisual transfer modes (voice-over and audio description) in order to research alternative working flows that may guarantee higher accessibility levels. Although limited in scope due to funding restrictions and its national scope, it is an innovation in audiovisual translation because until now research on machine translation in this field has mainly dealt with subtitling. ALST will hopefully be the first step in the application of such technologies in both voice-over and audio description and open new research horizons at international level.

António Branco and Petya Osenova: QTLeap – Quality Translation with Deep Language Engineering Approaches

The goal of this project is to contribute for the advancement of quality MT by pursuing an approach that further relies on semantics and opens the way to higher quality translation. We build on the complementarity of the two pillars of language technology — symbolic and probabilistic — and seek to advance their hybridization. We explore combinations of them that amplify their strengths and mitigate their drawbacks, along the development of three MT pilot systems that progressively seek to integrate deep language engineering approaches. The construction of deep treebanks has progressed to be delivering now the first significant Parallel DeepBanks, where pairs of synonymous sentences from different languages are annotated with their fully-fledged gram-

matical representations, up to the level of their semantic representation. The construction of Linked Open Data and other semantic resources, in turn, has progressed now to support impactful application of lexical semantic processing that handles and resolves referential and conceptual ambiguity. These cutting edge advances permit for the cross-lingual alignment supporting translation to be established at the level of deeper semantic representation.

Fatiha Sadat: The ASMAT project – Arabic Social Media Analysis Tools

The main objective of the ASMAT project is to make available a comprehensive set of language resources and tools covering Arabic dialects in social media context. Current Arabic NLP tools are capable of analysing large part of standard Arabic, but fail short of handling the dialects and the social media domain. To this end, the project aims to create tools for Arabic language and its varieties following certain tasks: language and dialect identification; dialect to standard (MSA) mapping and vice versa; automatic machine translation from any Arabic dialect to English and French. More specifically, the ASMAT project deals with the Maghrebi (North African) Arabic dialects for machine translation with very scarce resources.

Marko Tadić: XLike – Cross-Lingual Knowledge Extraction

The goal of the XLike project is to develop technology to monitor and aggregate knowledge that is currently spread across mainstream and social media, and to enable cross-lingual services for publishers, media monitoring and business intelligence. The aim is to combine scientific insights from several scientific areas to contribute in the area of cross-lingual text understanding. By combining modern computational linguistics, machine translation, machine learning, text mining and semantic technologies we plan to deal with the following two key open research problems: (1) to extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases, and; (2) to adapt linguistic tech-

niques and crowdsourcing to deal with irregularities in informal language used primarily in social media. The developed technology will be language-agnostic, while within the project we specifically address English, German, Spanish, Chinese as major world languages and Catalan, Slovenian and Croatian as minority languages. Knowledge resources from Linked Open Data cloud (e.g. Wikipedia, DBpedia, Wordnets etc.) will be used with special focus on general common sense knowledge base CycKB, that will be used as Interlingua. A number of different methods to translate from natural language to the selected formal language that serves as our Interlingua are being explored, among others also SMT.

Antonio Toral, Tommi Pirinen, Andy Way, Raphaël Rubino, Gema Ramírez-Sánchez, Sergio Ortiz, Miquel Esplà, Felipe Sánchez-Martínez, Mikel Forcada, Nikola Ljubešić, Prokopis Prokopidis and Vasilis Papavasiliou: Abu-MaTran: Automatic building of Machine Translation

Abu-MaTran seeks to enhance industry–academia cooperation as a key aspect to tackle one of Europe's biggest challenges: multilingualism. We aim to increase the hitherto low industrial adoption of machine translation by identifying crucial cutting-edge research techniques (automatic acquisition of corpora and linguistic resources, pivot-language techniques, linguistically augmented statistical translation and diagnostic evaluation), making them suitable for commercial exploitation. We also aim to transfer back to academia the know-how of industry to make research results more robust. We work on a case study of strategic interest for Europe: machine translation for the language of a new member state (Croatian) and related languages. All the resources produced will be released as free/open-source software, resulting in effective knowledge transfer beyond the consortium. The project has a strong emphasis on dissemination, through the organisation of workshops that focus on inter-sectoral knowledge transfer. Finally, we have a comprehensive outreach plan, including the establishment of a Linguistic Olympiad in Spain, open-day activities and the participation in the Google Summer of Code. At EAMT 2014 we

will present the results of the first milestone of the project (July 2013), a general-domain MT system for English–Croatian based on free/open-source software and publicly available data, released on July 1st 2013 to mark Croatia's accession to the EU. We will also present ongoing work towards the second milestone (December 2014) including (i) a domain-specific MT system for English–Croatian in the domain of tourism, (ii) generation of synthetic English–Croatian data via Slovene using quality estimation, (iii) outcomes of the first edition of the Linguistic Olympiad of Spain (September 2013–March 2014), etc.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris and Hans Uszkoreit: QTLaunchPad

QTLaunchPad is dedicated to overcoming quality barriers in machine and human translation and in language technologies. It is preparing for a large-scale translation quality initiative for Europe. One of the key contributions of QTLaunchPad is the Multidimensional Quality Metrics (MQM), a customizable system that provides analytic methods to assess machine translation output. This system has been used to assess results from top-performing WMT systems and customer data provided by language service providers. Analysis in the project has focused on "almost good" translations, those segments where MT systems produce results that can be easily fixed, to understand the barriers that impact the best MT systems. It has also worked on the development of quality estimation and linguistic evaluation techniques to assist MT processes to identify those segments that are good enough to use as is, those that can be easily repaired by human post-editors, and those that should be discarded and translated from scratch. Key findings include the identification of linguistic structures in English that are particularly likely to trigger problems for MT systems of different types (e.g., use of *-ing* verb forms, nongenitive uses of *of*, and differences across languages in permissible positions within sentences). These findings were only possible by combining the insights of human evaluators and the output of computational tools. The insights gained from this analysis

will be of use to developers seeking to improve MT systems and to implementers seeking to integrate MT into "real world" production chains that include MT, human translators or posteditors, and other technologies.

Vincent Vandeghinste and Ineke Schuurman: Able-to-Include: Improving Accessibility for people with Intellectual Disabilities

While this project has the wider goal to improve the accessibility of the information society for people with intellectual disabilities, one of the important means for achieving this is the automated text-to-pictogram translator which has been developed for Dutch (webservices.ccl.kuleuven.be/picto/). In the Able-to-include project we localize the text-to-pictogram translator in order to make it work for Spanish and English, besides Dutch. Pictograms have been linked to Wordnet synsets from the Dutch lexical-semantic database Cornetto (Vandeghinste & Schuurman @ LREC2014). We will establish links between Princeton Wordnet and the pictograms by using the equivalence relations which link Cornetto synsets to Princeton synsets. In a second stage we will establish the link between the pictograms and the Spanish Wordnet by using the equivalence relations that are provided between the Spanish Wordnet and Princeton Wordnet. In order to obtain a full localisation of the Text2Picto translator, we will also have to adapt the linguistic components to English and Spanish. We will also provide Picto2Text which translates a sequence of pictograms into natural language, and which will be used in combination with a pictogram-selection mechanism that serves as input method for people with writing difficulties. The tools work with two different pictogram sets, Beta and Sclera. These tools will be used in several pilot projects involving actual user organisations and users with intellectual disabilities, allowing to measure their impact on the daily lives of the target group.

Vincent Vandeghinste, Tom Vanallemeersch, Véronique Hoste, Marie-Francine Moens, Patrick Wambacq, Karin Coninx and Ken De Wachter: Smart Computer Aided Translation Environment

In the SCATE project we aim at improving the translators' efficiency. Commercial translation tools are faced with ever higher productivity requirements imposed by the globalisation of business activities and the increasing information flow. The SCATE project intends to improve translators' efficiency along the following axes: exploitation of already translated data, translation evaluation, terminology extraction, speech recognition, workflows and personalised user interfaces.

User track
Poster Session B: Tuesday, 2014-06-17, 11:30–13:00

Sharon O'Brien, Joss Moorkens and Joris Vreeke: Kanjingo – A Mobile App for Post-Editing

This paper describes the Kanjingo post-editing application for smartphones. The application was developed using an agile methodology at the Centre for Global Intelligent Content (CNGL) at DCU and a first stage of user testing was conducted using content from Translators Without Borders. Initial feedback on this app was quite positive. Users identified some particular challenges, e.g. input and sensitivity limitations, insufficient Help, lack of automatic punctuation and capitalization. Development and further testing are ongoing and may include interactive MT, speech as input and focus on Asian languages as target languages in the future.

Federico Fancellu and Andy Way: Standard language variety conversion using SMT

Translation between varieties of the same language is a widespread reality in the localisation industry. However, monolingual SMT is still a solution that has not yet been adequately explored; to the best of our knowledge, previous work in this area has never directly applied SMT to varieties of the same language for the precise purpose of reducing the time and cost of human translation and editing of content that needs to be localised. In this paper, we start exploring the problem by deploying SMT to translate Brazilian Portuguese into European Portuguese. Our exploration is mainly based on the use of bilingual dictionaries to guide the decoder and modify the translation output. We also consider the option of mining a bilingual dictionary from word alignments obtained after standard SMT training. On good-quality data provided by Intel, we show that the SMT baseline already constitutes a strong system which in a number of experiments we fail to improve upon. We conjecture that bilingual dictionaries mined from client data would help if more heterogeneous training data were to be added.

Vicent Alabau and Luis A. Leiva: Collaborative Web UI Localization, or How to Build Feature-rich Multilingual Datasets

We present a method to generate feature rich multilingual parallel datasets for machine translation systems, including e.g. type of widget, user's locale, or geolocation. To support this argument, we have developed a bookmarklet that instruments arbitrary websites so that casual end users can modify their texts on demand. After surveying 52 users, we conclude that people is leaned toward using this method in lieu of other comparable alternatives. We validate our prototype in a controlled study with 10 users, showing that language resources can be easily generated.

Johann Roturier, David Silva and Linda Mitchell: Using the ACCEPT framework to conduct an online community-based translation evaluation study

This paper presents how a novel evaluation framework was used to collect translation ratings thanks to users of an online German-speaking support community in the IT domain. Using an innovative data collection approach and mechanism, this paper shows that segment-level ratings can be collected in an effective manner. The collection mechanism leverages the ACCEPT evaluation framework which allows data collection to be triggered from online environments in which community users interact on a regular basis.

Mark Fishel and Rico Sennrich: Handling Technical OOVs in SMT

We present a project on machine translation of software help desk tickets, a highly technical text domain. The main source of translation errors were out-of-vocabulary tokens (OOVs), most of which were either in-domain German compounds or technical token sequences that must be preserved verbatim in the output. We describe our efforts on compound splitting and treatment of non-translatable tokens, which lead to a significant translation quality gain.

Research track
Tuesday, 2014-06-17, 14:30–15:30

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis and Hans Uszkoreit: Using a new analytic measure for the annotation and analysis of MT errors on real data

This work presents the new flexible Multidimensional Quality Metrics (MQM) framework and uses it to analyze the performance of state-of-the-art machine translation systems, focusing on "nearly

acceptable" translated sentences. A selection of WMT news data and "customer" data provided by language service providers (LSPs) in four language pairs was annotated using MQM issue types and examined in terms of the types of errors found in it. Despite criticisms of WMT data by the LSPs, an examination of the resulting errors and patterns for both types of data shows that they are strikingly consistent, with more variation between language pairs and system types than between text types. These results validate the use of WMT data in an analytic approach to assessing quality and show that analytic approaches represent a useful addition to more traditional assessment methodologies such as BLEU or METEOR.

Nicholas Ruiz and Marcello Federico: Complexity of Spoken Versus Written Language for Machine Translation

When machine translation researchers participate in evaluation tasks, they typically design their primary submissions using ideas that are not genre-specific. In fact, their systems look much the same from one evaluation campaign to another. In this paper, we analyze two popular genres: spoken language and written news, using publicly available corpora which stem from the popular WMT and IWSLT evaluation campaigns. We show that there is a sufficient amount of difference between the two genres that particular statistical modeling strategies should be applied to each task. We identify translation problems that are unique to each translation task and advise researchers of these phenomena to focus their efforts on the particular task.

Research track
Wednesday, 2014-06-18, 9:00–10:30

Sheila Castilho, Sharon O'Brien and Fabio Alves: Does post-editing increase usability? A study with Brazilian Portuguese as Target Language

It is often assumed that raw MT output requires post-editing if it is to be used for more than gisting purposes. However, we know little about how end users engage with raw machine translated text or post-edited text, or how usable this text is, in particular if users have to follow instructions and act on them. The research project described here measures the usability of raw machine translated text for Brazilian Portuguese as a target language and compares that with a post-edited version of the text. Two groups of 9 users each used either the raw MT or the post-edited version and carried out tasks using a PC-based security product. Usability was measured using an eye tracker and cognitive, temporal and pragmatic measures of usability, and satisfaction was measured using a post-task questionnaire. Results indicate that post-editing significantly increases the usability of machine translated text.

Maja Popović, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis and Hans Uszkoreit: Relations between different types of post-editing operations, cognitive effort and temporal effort

Despite the growing interest in and use of machine translation post-edited outputs, there is little research work exploring different types of post-editing operations, i.e. types of translation errors corrected by post-editing. This work investigates five types of post-edit operations and their relation with cognitive post-editing effort (quality level) and postediting time. Our results show that for French-to-English and English-to-Spanish translation outputs, lexical and word order edit operations require most cognitive effort, lexical edits require most time, whereas removing additions has a low impact both on quality and on time. It is also shown that the sentence length is an important factor for the post-editing time.

Oliver Čulo: Approaching Machine Translation from Translation Studies – a perspective on commonalities, potentials, differences

The exchange between Translation Studies (TS) and Machine Translation (MT) has been relatively rare. However, given recent developments in both fields like increased importance of post-editing and reintegration of linguistic and translational knowledge into hybrid systems, it seems desirable to intensify the exchange. This paper aims to contribute to bridging the gap between the two fields. I give a brief account of the changing perspective of TS scholars on the field of translation as a whole, including MT, leading to a more open concept of translation. I also point out some potential for knowledge transfer from TS to MT, the idea here centring around the adoption of text-centric notions from TS both for the further development of MT systems and the study of post-editing phenomena. The paper concludes by suggesting further steps to be taken in order to facilitate an intensified future exchange.

User track
Wednesday, 2014-06-18, 10:30–11:00 and 11:30–12:30

Raivis Skadiņš, Inguna Skadiņa, Mārcis Pinnis, Andrejs Vasiļjevs and Tomáš Hudík: Application of Machine Translation in Localization into low-resourced languages

This paper evaluates the impact of machine translation on the software localization process and the daily work of professional translators when SMT is applied to low-resourced languages with rich morphology. Translation from English into six low-resourced languages (Czech, Estonian, Hungarian, Latvian, Lithuanian and Polish) from different language groups are examined. Quality, usability and applicability of SMT for professional translation were evaluated. The building of domain and project tailored SMT systems for localization purposes was evaluated in two setups. The results

of the first evaluation were used to improve SMT systems and MT platform. The second evaluation analysed a more complex situation considering tag translation and its effects on the translator's productivity.

Marcin Junczys-Dowmunt and Bruno Pouliquen: SMT of German Patents at ORGNAME: Decompounding and Verb Structure Pre-reordering

We describe fragments of the SMT pipeline at WIPO for German as a source language. Two subsystems are discussed in detail: word decompounding and verb structure pre-reordering. Apart from automatic evaluation results for both subsystems, for the pre-reordering mechanism manual evaluation results are reported.

Antonio Toral, Raphael Rubino, Miquel Esplà, Tommi Pirinen, Andy Way and Gema Ramírez-Sánchez: Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain

We present an extrinsic evaluation of crawlers of parallel corpora from multilingual web sites in machine translation (MT). Our case study is on Croatian to English translation in the tourism domain. Given two crawlers, we build phrase-based statistical MT systems on the datasets produced by each crawler using different settings. We also combine the best datasets produced by each crawler (union and intersection) to build additional MT systems. Finally we combine the best of the previous systems (union) with general-domain data. This last system outperforms all the previous systems built on crawled data as well as two baselines (a system built on general-domain data and a well known online MT system).

About Sponsors

Bloomberg

Bloomberg technology helps drive the world's financial markets. We provide communications platforms, data, analytics, trading platforms, news and information for the world's leading financial market participants. We deliver through our unrivalled software, digital platforms, mobile applications and state of the art hardware developed by Bloomberg technologists for Bloomberg customers. Our over 3,000 technologists work to define, architect, build and deploy complete systems and solutions that anticipate and fulfil our clients' needs and market demands. We offer critical enterprise computing solutions for the financial services industry to help organizations deliver, decipher and manage data to meet their organizational needs and growing regulatory requirements. As well, we are known for providing unique biometric security to our customers for over a decade enabling them to access their service securely from a desktop computer, mobile phone or tablet. We have provided communications platforms, true electronic social networking and workflow connectivity in the financial world for over two decades.

European Language Resources Association

ELRA's missions are to promote language resources for the Human Language Technology (HLT) sector, and to evaluate language engineering technologies. To achieve these two major missions, we offer a range of services, listed below and described in the "Services around Language Resources" section: a) Identification of language resources; b) Promotion of the production of language resources; c) Production of language resources; d) Validation of language resources; e) Evaluation of systems, products, tools, etc., related to language resources; f) Distribution of language resources; g) Standardisation. The promotion of the production of language resources also includes our support of the infrastructure for evaluation campaigns and our support in developing a scientific field of language resources and evaluation.

Springer

Our business is publishing. Throughout the world, we provide scientific and professional communities with superior specialist information – produced by authors and colleagues across cultures in a nurtured collegial atmosphere of which we are justifiably proud. We foster communication among our customers – researchers, students and professionals – enabling them to work more efficiently, thereby advancing knowledge and learning. Our dynamic growth allows us to invest continually all over the world. We think ahead, move fast and promote change: creative business models, inventive products, and mutually beneficial international partnerships have established us as a trusted supplier and pioneer in the information age.

ISBN 978-953-55375-4-0

Conference Programme and Book of Abstracts
17th Annual Conference of the European Association for Machine Translation
EAMT2014

Dubrovnik, Croatia, 16th-18th June 2014

Sponsored by

Bloomberg

EUROPEAN ASSOCIATION ELRA LANGUAGE RESOURCES

Springer

ISBN 978-953-55375-4-0