Baltic J. Modern Computing, Vol. 4 (2016), No. 2, 376-400

Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)

Riga, Latvia, 2016

Projects/Products

TectoMT – a Deep-Linguistic Core of the Combined Chimera MT system

Martin POPEL, Roman SUDARIKOV, Ondřej BOJAR, Rudolf ROSA, Jan HAJIČ

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské nám. 25, CZ-11800 Prague 1, Czech Republic

{popel,sudarikov,bojar,rosa,hajic}@ufal.mff.cuni.cz

Abstract. Chimera is a machine translation system that combines the TectoMT deep-linguistic core with phrase-based MT system Moses. For English–Czech pair it also uses the Depfix post-correction system. All the components run on Unix/Linux platform and are open source (available from Perl repository CPAN and the LINDAT/CLARIN repository). The main website is https://ufal.mff.cuni.cz/tectomt. The development is currently supported by the QTLeap 7th FP project (http://qtleap.eu).

TectoMT and Chimera

TectoMT (the deep-linguistic core of Chimera) is an open-source MT system based on the Treex platform for general natural-language processing. TectoMT uses a combination of rule-based and statistical (trained) modules ("blocks" in Treex terminology), with a statistical transfer based on HMTM (Hidden Markov Tree Model) at the level of a deep, so-called tectogrammatical representation of sentence structure. In the Chimera combination, TectoMT is complemented by a Moses PB-SMT system (factored setup with additional language models over morphological tags) and optionally also by an automatic postprocessing (correction) component called Depfix. Chimera can be thus characterized as a hybrid system that combines statistical MT with deep linguistic analysis and automatic post-correction system, which is useful especially for translation into inflectionally rich languages. The three systems are combined serially: TectoMT runs first, then an additional Moses phrase table is extracted from TectoMT's input and output. The additional table is then used in a weighted combination with a large Moses translation table to produce pre-final output. Depfix then re-parses the output (as well as input) and generates the final output based on rules reflecting morphosyntactic properties of the target language.

Chimera was transferred from English–Czech to additional three language pairs (English to Dutch, Portuguese and Spanish) within the QTLeap 7th EU project.

References

Dušek, O., Gomes, L., Novák, M., Popel, M., Rosa, R. (2015). New Language Pairs in TectoMT. *Proceedings of the 10th Workshop on Machine Translation*, ISBN 978-1-941643-32-7, ACL, Stroudsburg, PA, USA, 98–104.

Rosa, R., Dušek, O., Novák, M., Popel. M. (2015). Translation Model Interpolation for Domain Adaptation in TectoMT. *Proceedings of the 1st Deep Machine Translation Workshop*, ISBN 978-80-904571-7-1, Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic, 89–96.

Bojar, O., Tamchyna, A. (2015). CUNI in WMT15: Chimera Strikes Again. *Proceedings of the 10th Workshop on Machine Translation*, ISBN 978-1-941643-32-7, ACL, Stroudsburg, PA, USA, 79–83.

WiTKoM - Virtual Sign Language Translator Project

Katarzyna BARCZEWSKA[1], Jakub GAŁKA[1,2], Filip MALAWSKI[1], Mariusz MĄSIOR[1], Dorota SZULC[1], Tomasz WILCZYŃSKI[1,2], Krzysztof WRÓBEL[1]

[1]AGH University of Science and Technology, Department of Electronics, Poland
[2]VoicePIN.com Sp. z o. o., Poland

jgalka@agh.edu.pl

Abstract. WiTKoM (Virtual Sign Language Translator) is an interdisciplinary research project carried out by AGH University of Science and Technology and VoicePIN.com, which aims to create a Polish Sign Language (PJM, pl. Polski Język Migowy) translator. This work was supported by the Polish National Centre for Research and Development – Applied Research Program under Grant PBS2/B3/21/2013 tilted Virtual sign language translator. Website: www.witkom.info.

Description

WiTKoM is intended to promote the social inclusion. Hearing-impaired people constitute a considerable language minority in Poland. PJM is currently experiencing a renaissance, having 50,000–100,000 users, according to recent statistics. The practical goal of the project is to develop the technology and solutions for communication support for hearing-impaired.

Project WiTKoM consists of various separate elements, responsible for the particular stage of Polish to PJM translation, and statement creation by the signing avatar. Within the project, several developments have been made: know-how, technologies, and automatic PJM translation software. The developed technology includes:

- real-time user-independent gesture recognition system;
- PL–PJM translator, from Polish sentence analysis, through machine translation methods, to avatar transcription system employing HamNoSys;
- mobile application prototype for gesture recognition;
- accelerometer-based sensor glove for gesture motion acquisition;
- computer application for conducting automatic dialogues, module for building and managing dialogue;
- computer application for multi-stream data acquisition and management;
- a rich set of developing tools for conducting researches in the fields of image processing and gesture recognition;
- annotated PJM gesture corpus acquired with RGB and depth cameras;
- Polish-PJM parallel corpus for machine translation research.

WiTKoM project is still in progress. The main focus is currently on the development of continuous statement recognition algorithms and practical use-case deployments.

Multi-Level Quality Prediction with QuEst++

Gustavo H. PAETZOLD, Lucia SPECIA

University of Sheffield, Department of Computer Science
Western Bank, Sheffield, South Yorkshire S10 2TN, United Kingdom

ghpaetzold1@sheffield.ac.uk, l.specia@sheffield.ac.uk

Abstract. We introduce QuEst++: an open-source multi-level Machine Translation Quality Estimation framework. The core of the framework is implemented in Java, and wrappers for a Machine Learning module implemented in Python are provided. QuEst++'s development was funded by the EAMT, as well as by the QT21 and EXPERT projects. The framework is distributed under the BSD license and can be downloaded from: https://github.com/ghpaetzold/questplusplus.

Description

QuEst++ is an extended and improved version of QuEst, a framework for Machine Translation Quality Estimation (Specia et. al., 2013). While the original QuEst framework provides only sentence-level Quality Estimation, QuEst++ is a multi-level framework that combines solutions to word-, sentence- and document-level Quality Estimation in the same pipeline.

QuEst++ is composed of two modules: Feature Extraction and Machine Learning. The Feature Extraction module provides access to 43 word-level features, 148 sentence-level features and 67 document-level features. If certain resources needed for feature extraction are not provided, QuEst++ employs its Automatic Resource Generation routines to produce the resources.

The Machine Learning module offers all the utilities previously included in QuEst, and also an easy-to-use interface to CRFSuite (Okazaki, 2007), which allows for state-of-the-art Conditional Random Field models to be trained. As reported in the WMT 2015 Quality Estimation tasks' results (Bojar et al., 2015), QuEst++ itself performs well as compared to other systems. In addition, it has been used as the basis to create more advanced approaches. More details on QuEst++ can be found in Specia et al. (2015).

References

Specia, L., Paetzold, G., Scarton, C. (2015). Multi-level Translation Quality Prediction with QuEst++. In *ACL-IJCNLP 2015 System Demonstrations*, 115–120, Beijing.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C. Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, S., Specia, L., Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *WMT* , 1–46, Lisbon.

Specia, L., Shah, K., De Souza, J. G., & Cohn, T. (2013). QuEst-A Translation Quality Estimation Framework. In ACL 2013, 79-84, Sofia.

Okazaki, N. (2007). CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs). URL http://www. chokkan. org/software/crfsuite.

Apertium: A Free/Open-Source Platform for Machine Translation and Basic Language Technology

Mikel L. FORCADA[1] and Francis M. TYERS[2]

[1] Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Alacant
[2] HSL-fakultehta, UiT Norgga árktalas universitehta, Romsa

mlf@dlsi.ua.es, francis.tyers@uit.no

Abstract. Apertium is a free/open-source platform for rule-based machine translation which was started in 2005. The Apertium community is using it to build machine translation systems for a variety of language pairs, especially, but not only, for related-language pairs, where shallow transfer suffices to produce good quality translations. Apertium is also being used to develop monolingual language processors for these languages. We present the Apertium platform: the translation engine, the encoding of linguistic data, and the tools developed around the platform.

Description

Apertium (wiki.apertium.org, Forcada et al. 2011) is a free/open-source platform for machine translation (MT), providing a shallow-transfer rule-based MT engine, data (dictionaries, rules) for more than 40 language pairs, and a wide variety of software to manage, convert and exploit the data and run the engine in a variety of environments. Apertium is also the free/open-source community developing it since 2005.

Apertium is built as a pipeline in which the source text is gradually converted into target text. The pipeline contains, among others, filters to handle markup, finite-state morphological analysers and generators using both a native format and the Helsinki Finite-State Toolkit (HFST) one, part-of-speech taggers using Constraint Grammar, hidden Markov Models, and sliding-window classifiers, and shallow-transfer rules capable of identifying, transforming and reordering syntactic chunks. These components may also be used for other human-language technologies besides MT.

Apertium is also available in several end-user products: apps for Android smartphones and tablets, a plugin for the OmegaT computer-aided translation environment, and a number of stand-alone applications for desktops. It also powers the multilingual open-content management environment, Wikimedia Content Translation, used to translate Wikipedia articles.

Finally, Apertium is both a research and a business platform, enabling reproducibility and transferability of results and successful marketing of MT products and services.

References

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. and Tyers, F. M. (2011) Apertium: a free/open-source platform for rule-based machine translation. Machine Translation 24(1) 1–18

BabelDr: A Web Platform for Rapid Construction of Phrasebook-Style Medical Speech Translation Applications

Pierrette BOUILLON, Hervé SPECHBACH

FTI/TIM, Geneva University, Geneva, Switzerland
Hôpitaux Universitaires de Genève, Geneva, Switzerland

Pierrette.Bouillon@unige.ch, Herve.Spechbach@hcuge.ch

Abstract. BabelDr (http://babeldr.unige.ch/) is a joint project of Geneva's Faculty of Translation and Interpretation (FTI) and University Hospitals (HUG), that has been active since July 2015. The goal is to develop methods that allow rapid prototyping of medium-vocabulary web-enabled medical speech translators, with particular emphasis on languages spoken by victims of the current European refugee crisis. A demonstrator system freely available on the project site translates spoken French medical examination questions into four languages.

BabelDr (http://babeldr.unige.ch/) is a joint project of Geneva University's Faculty of Translation and Interpretation (FTI/TIM) and Geneva University Hospital (HUG), active since July 2015 under funding from "La fondation privée des HUG". The goal is to develop methods that allow rapid prototyping of medium-vocabulary web-enabled medical speech translators, with particular emphasis on languages spoken by migrants. The application can be characterised as a flexible speech-enabled phrasebook (Rayner et al 2015). Semantic coverage consists of a prespecified set of utterance-types (around 2000 in the current version), but users can use a wide variety of surface forms when speaking to the system. Each utterance-type is associated with a canonical source-language version, which is rendered into the target languages by suitably qualified translation experts. The central design goals are to ensure that a) translations are completely reliable, b) the bulk of the work can be performed directly by translation experts, with minimal or no involvement from language engineers, c) speech recognition performance is excellent for in-coverage data and adequate even for new users, d) new versions of the live app can be quickly deployed over the web, enabling rapid updating of coverage in response to requests from medical staff, e) new target languages can easily be added, enabling flexibility in the face of changing patient demographics. A demonstrator system freely accessible at http://babeldr.unige.ch/demos-and-resources/ translates French into Spanish, Italian, Arabic and Tigrinya.

References

Rayner, M. et al. (2015). Helping Domain Experts Build Phrasal Speech Translation Systems. Proceedings of the Workshop on Future and Emerging Trends in Language Technology, Sevilla, Spain, pages 1-12.

SCATE – Smart Computer Aided Translation Environment

V. VANDEGHINSTE[1], T. VANALLEMEERSCH[1], L. AUGUSTINUS[1], J. PELEMANS[1], G. HEYMANS[1], I. VAN DER LEK-CIUDIN[1], A. TEZCAN[2], D. DEGRAEN[3], J. VAN DEN BERGH[3], L. MACKEN[2], E. LEFEVER[2], M. MOENS[1], P. WAMBACQ[1], F. STEURS[1], K. CONINX[3], F. VAN EYNDE[1]

[1]University of Leuven – Departments of Linguistics, Computer Science, and Electronical Engineering; [2]Ghent University – LT3; [3]Hasselt University - EDM

scate-board@ccl.kuleuven.be

Abstract. The SCATE project aims at improving translators' efficiency through improvements in translation technology, evaluation of computer-aided translation, terminology extraction from comparable corpora, speech recognition accuracy, and work flows and personalised user interfaces. It is funded by IWT-SBO*, project nr. 130041. http://www.ccl.kuleuven.be/scate/

Envisaged Project Results

We present the envisaged results of SCATE, now the project is mid-term, with two more years to go.

We have surveyed and observed translators with respect to the following aspects: human-machine interaction in post-editing, human acquisition of domain knowledge and terminology, and workflow usage and interface personalization. We are researching different computer-aided translation (CAT) technologies, such as syntax-based fuzzy matching and concordancing, tools for speedier and more consistent collaborative translation, automated term extraction methods from comparable corpora, and integrated models and domain adaptation for speech as a post-editing method. Concerning MT Technology, we are working on syntax-based transduction, taxonomy-based confidence estimation metrics and speech translation. For these purposes, we have developed the following resources: a taxonomy of MT errors and manually annotated corpus of MT errors. Concerning the user interface, we are developing new approaches towards visualisation of translation features and towards flexible user interfaces. By the end of the project, we intend to integrate most of these aspects in a demonstration system that translates from English to Dutch.

We are interested in feedback from language service providers and translators: what do you consider useful and interesting – how can we improve your translation environment?

* The Flemish Agency for Innovation through Science and Technology, Strategic Basic Research.

HimL: Health in My Language

Barry HADDOW, Alex FRASER

Ludwig Maximilian University of Munich

fraser@cis.lmu.de

Abstract. HimL (www.himl.eu) is a three-year EU H2020 innovation action, which started in February 2015. Its aim is to increase the availability of public health information via automatic translation. Targeting languages of Central and Eastern Europe (Czech, German, Polish and Romanian) we aim to produce translations which are adapted to the health domain, semantically accurate and morphologically correct. The project is coordinated by Barry Haddow (University of Edinburgh) and includes two additional academic partners (Charles University and LMU Munich), one integration partner (Lingea) and two user partners (NHS 24 and Cochrane).

Description

In HimL we aim to deploy and evaluate machine translation systems for the public health domain, addressing domain adaptation, semantic accuracy and target morphology. The systems are used to tramslate content for NHS 24 (Scotland's national telehealth organisation) and Cochrane (an international NGO that produces systematic reviews of healthcare topics). The project has now been running for over a year, and we have already developed the first release of our translation systems and used them to translate the user partner websites. To build these systems, we have collected a large and diverse training set and are analysing the performance of existing domain adaptation techniques in combining these resources, as well as investigating the use of neural models in domain adaptation. We have been developing our corrective approaches to morphology handling, using machine learning to provide language independence, as well as extending the two-step approach to morphology to handle a wider range of phenomena, and new language pairs. For improved semantic accuracy we have experimented with using semantic roles to make sure important information is not lost, as well as developing methods to remove semantically incorrect translations from the model, and analysing the problems that arise in the translation of negation. The goal of semantic accuracy is supported by our development of new human and automatic semantic evaluation measures based on the UCCA (universal conceptual cognitive annotation) framework.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.

OPUS – Parallel Corpora for Everyone

Jörg TIEDEMANN

Department of Modern Languages, University of Helsinki, 00014 Helsinki, Finland

jorg.tiedemann@helsinki.fi

Abstract. Abstract. OPUS is a large collection of freely available parallel corpora that we provide in various formats and packages. All data sets are completely aligned at the sentence level for all possible language pairs. OPUS covers over 200 languages and language variants with a total of about 3.2 billion sentences and sentence fragments containing over 28 billion tokens. The collection contains data from various sources and domains and each sub-corpus is provided in common data formats to make it easy to integrate them in research and development. OPUS also provides tools and on-line interfaces for exploring parts of the collection and is continuously growing in terms of size and coverage.

Description

Parallel data sets in OPUS[1] are freely available and cover various domains. The largest collections (in terms of volume) come from political and administrative sources such as the European Commission and user-provided movie subtitles in various languages. Other sources include software localisation, multilingual news providers, translated descriptions of medical products, religious texts and multilingual wikis and other websites. OPUS is organised by source and one of the main principles of the collection is to preserve the original data structures (file structure, formatting, meta-data) as much as possible. The goal of our project is to make the collection applicable as widely as possible. Currently, OPUS comprises 3.2 billion sentences with over 28 billion tokens in total. An important principle is complete sentence alignment for all language pairs, thus, supporting even low-density languages and unusual combinations. There are bitexts such as Arabic–Korean or Indonesian–Latvian with over one million translation units among the 12,572 language pairs with their 10.8 billion translation units in total. We provide the data in standalone XML and stand-off alignment (as its native format) but also commonly used formats such as TMX and aligned plain Unicode text. The latest edition of OPUS contains parallel sentences extracted from Wikipedia and a significantly extended collection of movie subtitles, now also including intra-lingual alignments of alternative translations. Furthermore, we also provide word alignments and phrase-tables for statistical machine translation (SMT) ready to be used in common SMT toolboxes. A subset of our data is also available via on-line search interfaces. Feedback and contributions are welcome.

[1] http://opus.lingfil.uu.se

Integration of Machine Translation Paradigms

Marta R. COSTA-JUSSÀ

TALP Research Center, Universitat Politècnica de Catalunya, Barcelona

marta.ruiz@upc.edu

Abstract. Machine Translation (MT) is a highly interdisciplinary and multidisciplinary field since it is approached from the point of view of engineering, computer science, informatics, statistics and linguists. The goal of this research project is to approach the different profiles in the MT community by providing a new integrated MT paradigm which mainly includes linguistic technologies and statistical algorithms.

Description

The proposed new paradigm in this project provides solutions to current MT challenges such as unknown words, reordering and semantic ambiguities. The project focuses on three of the most spoken languages: Chinese, Spanish and English. These language pairs do not only involve many economic and cultural interests, but they also include some of the most relevant MT challenges such as morphological, syntactic and semantic variations. This project is funded under FP7-PEOPLE-2011-299251-IOF MarieCurie International Outgoing Fellowship[1]. The project duration is from 2012-12-07 to 2016-07-08. The project coordinator center is Universitat Politècnica de Catalunya (UPC, Barcelona) and the supervisor is Prof. José A. R. Fonollosa. The host institution has been the Institute for Infocomm Research (I2R, Singapore) and the corresponding supervisors were Prof. Haizhou Li and Dr. Rafael E. Banchs. The MarieCurie Researcher is Dr. Marta R. Costa-jussà. See a complete list of the project publications[2].

References

Costa-jussà, M. R. How much Hybridization does MT Need? *Journal of the Association for Information Science and Technology*, 6(10), 2015

Costa-jussà, M. R. and Centelles, J. Description of the Chinese-to-Spanish RBMT System Developed with a Hybrid Combination of Human Annotation and Statistical Techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(1), 2016

Costa-jussà, M. R and Fonollosa, J.A.R. Character-based Neural MT Proc. of ACL, 2016

[1] http://www.costa-jussa.com/projects/ongoing/imtrap/

[2] http://www.costa-jussa.com

STAR Transit & STAR MT: Morphologically Generated Additional Information for Improving MT Quality

Nadira HOFMANN

STAR Group, Wiesholz 35, 8262 Ramsen, Switzerland

nadira.hofmann@star-group.net

Abstract. As an experienced developer of language processing solutions, STAR Group applies its proven technologies to MT training: Transit's morphological support for 80+ languages is used to extract inflected terminology from dictionaries and reference material without any additional effort. This additional input for the engine training noticeably improves the MT quality – especially in morphologically rich languages.

Description

For STAR's MT system STAR MT, validated customer-specific dictionaries are used for MT training. However, dictionaries contain terms in canonical forms, while text corpora predominantly contains inflected forms. A large proportion of the terminological potential would remain untapped if engines were only trained with canonical forms.

This is where STAR's experience pays off: The TMS Transit offers morphological terminology support for over 80 languages -- not just simple stemming, but using linguistic expertise mapped out in rules that have been tried and tested throughout years.

As a result, inflected forms in the source and target language of a bilingual body of text are reliably identified and used to enrich the training material. This additional information is particularly valuable for the quality of the MT because it is validated twice: Customer-validated canonical forms that are retrieved from the dictionary along with their inflected forms that are actually used in the validated translation memory.

With this approach, STAR MT benefits from Transit's existing and proven morphological technology to create terminology that offers added value without investing any time. In practice, this brings about significant improvements in quality: For a customer-specific German-French engine, an additional 17% of terminology was extracted and the BLEU score increased by 1.1 points. The translations that are created from this were clearly selected as preferred translations by the translators who carried out the manual evaluation of the sentence BLEU lists.

References

Pinnis, M. (2015). Dynamic Terminology Integration Methods in Statistical Machine Translation, Proceedings of the 18th Annual Conference of the European Association for Machine Translation, 89-96.

AltLang: an Automatic Converter between Varieties of English, Spanish, French and Portuguese

Gema RAMÍREZ-SÁNCHEZ

Prompsit Language Engineering, Av. de la Unversitat s/n, 03202, Elx, Alacant, Spain

gramirez@prompsit.com

Abstract. AltLang is a rule-based automatic converter for language varieties. It deals with differences in spelling, lexicon and local grammar along with numeric, style and punctuation conventions. It is available for varieties of English, Spanish, French and Portuguese. AltLang is based on the GNU GPL-based free/open-source technologies of the Apertium platform, and is offered as a service by Prompsit. It can be tested through an online application at www.altlang.net.

Description

AltLang is a rule-based automatic converter for language varieties. It aims at minimising the effort required to generate content in different varieties of the same language by automatically replacing controlled existing differences between them. AltLang can convert between American and British English, Canadian and European French, Latin American and European Spanish and Brazilian and European Portuguese with high accuracy. The technology behind AltLang comes from the free/open-source Apertium platform (Forcada et al. 2011). Data for dictionaries and rules have been semi-automatically adapted from two main sources: Apertium data and freely available or client-owned bilingual corpora. AltLang's knowledge grows through continuous contributions from clients and testers. AltLang gives priority to user-defined vocabulary and translation memories: vocabulary is input as a simple two-column spreadsheet and translation memories are supported through standard TMX files. AltLang reports basic statistics about the conversions (number of words and characters in source and target and number of substitution, shifts, insertions and deletions). It also shows the individual changes performed to a document with a convenient colouring. All these features allow the user to have full control over the behaviour of AltLang. It deals with a wide range of file formats including some interchange and localization ones (.xliff, .po). It can be used through an API, AltLang's web application and is available in other third-party tools such as the MateCat translation platform. Prompsit offers customisation services for AltLang to add new languages, new language varieties or domain-adaptation.

References

Forcada, M. L.; Ginestí-Rosell, M.; Nordfalk, J.; O'Regan, J.; Ortiz-Rojas, S.; Perez-Ortiz, J. A.; Sanchez-Martínez, F.; Ramírez.-Sánchez, G.; and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation; Machine Translation, 25(2):127–144.

iEMS – Interactive Experiment Management System for Machine Translation

Pēteris ŅIKIFOROVS

p@mak.id.lv

Abstract. Interactive Experiment Management System (Interactive EMS or iEMS) is an experiment management system with a graphical user interface for designing and running statistical machine translation experiments. It is written in JavaScript and runs in all modern desktop browsers with no installation. iEMS produces a script that can be used to train a complete machine translation engine from scratch. There is an optional backend that can be used to launch and configure virtual servers on the Amazon Web Services cloud for running experiments. It is an open-source project licensed under the Apache 2.0 license. The development of iEMS is supported by the European Association for Machine Translation (EAMT). Project website: https://github.com/pdonald/iems

Description

Training a statistical machine translation engine consists of many steps. There are experiment management systems like Moses EMS and eman that help manage all these steps. However, their configuration is stored in text files which makes it is difficult to visualize the order of execution and adding or modifying steps may require diving into application code.

iEMS is an interactive experiment management system with a graphical user interface that aims to make it easy to design and run machine translation experiments from scratch. The main objective of the project is to have very little friction to get started. It is a single-page application written in JavaScript intended to run in modern desktop browsers without any setup.

In the application, there are several predefined tools that represent various steps in machine translation training that can be dragged and dropped onto a design surface. They can be linked together to set the order of execution. Tools can be grouped which allows quick swapping one set of tools with another.

There is also a backend that allows the user to run their experiments. The experiments can be run on a local host via the secure shell SSH or a new virtual server can be launched to run them on the Amazon Web Services cloud with a single click. Both regular (only pay for what you use) and spot (cheaper but can be terminated at any time) types of virtual servers are supported. Docker, a tool for automating the deployment of software inside Linux containers, is used for provisioning which means it is not necessary to compile or install dependencies for machine translation tools and setting up the server for use takes very little time.

The project is still in development. Its development is supported by the European Association for Machine Translation.

KantanLQR: A Platform for Human Evaluation of Machine Translation Output to Drive Engine Rapid Improvement

Laura CASANELLAS LURI

KantanMT, Invent Building, Dublin City University, Glasnevin, Dublin 9, Ireland

Laurac@kantanmt.com, Tonyod@kantanmt.com

Abstract: KantanLQR is a quality review tool dedicated to making human evaluation of machine translation (MT) faster, seamless and more efficient. Its final goal is to dramatically reduce the time of engine re-training. It is a tool aimed to project managers (PMs) and reviewers. It is part of the KantanMT suite of tools and includes comprehensive error typology based on industry standards that can be customized to address the needs of each specific project. Some of the functionalities include automatic workflow and report creation, as well as sophisticated visuals produced in real time. KantanLQR is cloud-based and does not require any software instalments or license, only a monthly subscription. www.kantanmt.com.

Description

KantanLQR technology removes the need of static forms by offering an automated workflow that can be customized to the dynamic character of quality definition and the industry requirements of fast, simple and customizable tools. Up until now the industry would have used different forms depending on the type of evaluation required; for instance, evaluation of an output that will be used for gisting should focus on *usability* and does not require a complex form that includes error typology, whereas a project with the goal of producing full post-edited publishable content will require something more comprehensive. KantanLQR provides this type of flexibility. With the use of this platform, the linguistic skills of the reviewers assigned to the project are maximized, as they can focus on the review and post-editing of the MT segments without being overburdened with complex interfaces.

This tool is designed for both reviewers and PMs to automate the review workflows involved in improving the quality and translation fidelity of their customized KantanMT engines. It will reduce the time required to quality review the MT engines by as much as 50% and it will improve team collaboration and workflow management. KantanLQR is based on the concept that the best MT solution is the one that can be intelligently customized and improved. The gains are twofold: the quality of the MT output is checked and assessed following specific quality requirements that have been set up by the PM, and the linguistic edits and feedback can be used to further retrain the engine.

As this process is automated, it reduces time. Segments are distributed among reviewers who work using the error typology that have been customized for the project. Results are tracked in real time and progress is easily monitored. The outcome becomes a radiography of an engine at a glance, as the report containing data gathered from all reviewers involved is displayed in a highly visual manner.

PangeaMT v3 – Customise Your Own Machine Translation Environment

Alexandre HELLE, Manuel HERRANZ

Pangeanic BI-Europe SL, Av de las Cortes Valencianas, 26-5 Of. 107, 46015 Valencia, Spain

a.helle@pangeanic.com, m.herranz@pangeanic.com

Abstract. PangeaMT is the tool developed by Pangeanic in 2010 to automate translation processes. Pangeanic, together with the Computer Science Institute (ITI) of Valencia has completed the development of PangeaMT v3 platform as part of the Spanish national project "COR: FAST AND EFFICIENT TRANSLATION MANAGEMENT TOOL" with funding from the Spanish research organization CDTI, dependant from the Ministry of Economy. PangeaMT (http://pangeamt.com/) is based on Moses and it runs on GNU/Linux.

Customized Machine Translation

PangeaMT is a full machine translation environment. Pangeanic was the first LSP in the world to make commercial use of a Moses version with it as reported in the FP7 Euromatrixplus[1] project and earlier at AMTA (Yuste 2010). The platform is capable of processing large translation volumes and manage training datasets. It is ideal for companies, corporations and institutions that produce large amounts of documentation. It can be used to train engines with the client's own previously translated material and terminology, overcoming some of Moses limitations. Clients are free to create and retrain engines with new material. Thanks to its interfaces, the system can accept calls from external systems to translate digital and web content, or from CAT (Computer Assisted Translation) tools to improve translator's productivity.

PangeaMT has now reached version 3. New features include the use of monolingual data managed by the user to train different language models at will, further improvements in machine translation customisation and hybridisation. Version 3 also offers new interface connections, custom-built hybridisation techniques which can easily be expanded, and even a choice of translation engines (Moses, Apertium and third-party services) and algorithms if required for particular language pairs.

References

Yuste, E. et al. (2010) PangeaMT – putting standards to work... well. In Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas – AMTA 2010, Denver. Available at: amta2010.amtaweb.org/AMTA/papers/4-04-HerranzYusteEtal.pdf

[1] http://www.euromatrixplus.net/moses-decoder/

Interlingual Translation in the Grammatical Framework (GF)

Aarne RANTA, Krasimir ANGELOV, Thomas HALLGREN,
Prasanth KOLACHINA, Inari LISTENMAA

Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, 41296 Gothenburg, Sweden

{aarne,krasimir,hallgren,prasanth.kolachina,inari}@cse.gu.se

Abstract. Grammatical Framework (GF) is a grammar formalism for *multilingual grammars*, where many languages share a common *abstract syntax*. The abstract syntax can be used as a translation interlingua. GF was initially designed for domain-specific controlled language systems, but it has in recent years scaled up to wide-coverage translation as well. GF is an open-source project with a world-wide community that has created grammars for over 30 languages. Fifteen of these languages are currently included in a wide-coverage translator, which is available in the GF cloud and also as a mobile app, the GF Offline Translator, with a compact size that fits in a mobile phone. We plan to demo both kinds of systems in the conference.

Description

Interlingual grammar-based translation is probably as far from the current mainstream as an MT approach can be. However, it is a natural choice in controlled language applications, where a translator can work much like a compiler with multiple source and target languages and a shared abstract syntax. The original purpose of GF[1] was to make it easy to build such systems, by providing powerful engineering techniques and a general-purpose resource grammar library (RGL), which takes care of morphology and surface syntax.

Scaling up GF to wide-coverage translation is made possible by *layered interlinguas*, where deeper analysis levels (with narrower coverage) are backed up by shallower levels (with wider coverage). The interlingual grammar is combined with a *multilingual linked lexicon* (bootstrapped from open sources such as Wiktionary and Wordnet) and equipped with *statistical ranking* (based on treebanks) to select the best translations.[2,3]

[1] GF homepage: http://www.grammaticalframework.org/

[2] Available as GF cloud translator: http://cloud.grammaticalframework.org/wc.html,
an Android app: https://play.google.com/store/apps/details?id=org.grammaticalframework.ui.android
and an IOS app: https://itunes.apple.com/us/app/gf-offline-translator/id1023328422?mt=8

[3] This project is currently funded by REMU (Vetenskapsrådet): http://remu.grammaticalframework.org/

Domain-Specific Multilingual Translation for Producers of Information

Aarne RANTA, Krasimir ANGELOV, Markus FORSBERG, Thomas HALLGREN

Digital Grammars AB, Framnäsgatan 23, 41264, Gothenburg, Sweden

{aarne,krasimir,markus,thomas}@digitalgrammars.com

Abstract. Producers of multilingual information (webpages, manuals, etc) often need not solve the general problem of "translating anything", but can stay content with a more limited system, which however must be accurate. One way to do this is to define a *semantic interlingua*, which exactly fits the content to be translated. It has turned out that *algebraic datatypes*, and the way they are used in compilers to encode *abstract syntax*, are useful for this purpose. The grammar formalism Grammatical Framework (GF) was originally designed to support this kind of systems. While much of the current research in GF is focused on scaling it up to general-purpose translation, the creation of domain-specific systems has become a routine task, which is commercially exploited by the start-up company Digital Grammars AB. In the conference, we plan to show a demo of the tools for building such systems, as well as some actual systems built for customers.

Description

If we divide translation into the phases of *analysis* (of the source) and *generation* (of the target), we can identify some of the hardest problems as having to do with the analysis phase: in particular, *ambiguity* and *unexpected input*. The most radical solution to these problems is to *eliminate* them whenever possible. This is what *natural language generation* (NLG) is about: the source of documents in all target languages is *structured data*, such as databases or logical formulas. However, in real applications, such data is not always available in a clean form, but must be extracted from unstructured sources, typically from text. *Translation*, in this perspective, is NLG fed by parsing a source language. In GF, generation and parsing are defined simultaneously, since GF grammars are reversible.[1]

The mission of Digital Grammars AB[2] is to build high-quality documentation systems (either pure NLG or with translation) tailored for customers' needs. The technique enables both automatic and interactive translation, and is available for over 30 languages. Depending on the domain and on the work invested in building the system, we can reach publication quality (which is needed in real-time documentation) or come close to it (in which case we provide tools that support interaction; the system indicates confidence levels and displays translation alternatives with their semantic analyses). The product can be delivered in the form of batch translation jobs, as a web service, and as mobile (speech-enabled) applications.

[1] GF: http://www.grammaticalframework.org/

[2] Digital Grammars AB: http://www.digitalgrammars.com/

The EXPERT Project: Training the Future Experts in Translation Technology

Constantin ORĂSAN

Research Group in Computational Linguistics, University of Wolverhampton, Wulfruna St., Wolverhampton, WV1 1LY, United Kingdom

c.orasan@wlv.ac.uk

Abstract. The EXPERT project (http://expert-itn.eu) is an Initial Training Network (ITN) supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement no 317471. By appointing 15 fellows to work on related projects, the project aims to train the next generation of world-class researchers in the field of data-driven translation technology.

1. Description

This project presentation gives a brief overview of the EXPloiting Empirical appRoaches to Translation (EXPERT) project, an FP7 Marie Curie ITN in the field of translation technology. The purpose of the project is two-fold: to train 15 fellows to become future leaders in the field, by enabling them to pursue well-defined research projects, organising dedicated training events and enabling intersectoral and transnational secondments, and to advance the state of the art in data driven translation technologies.

The EXPERT project runs between 1st October 2012 and 30th September 2016, and is delivered by a consortium coordinated by University of Wolverhampton, UK and which contains five other academic partners: University of Malaga, Spain; University of Sheffield, UK; Saarland University, Germany; Dublin City University, Ireland and University of Amsterdam, Netherlands, as well as three industrial partners: Pangeanic, Spain; Translated, Italy and Hermes, Spain. In addition, the consortium benefits from the contribution of four associated partners: WordFast, France; Etrad, Argentina; Unbabel, Portugal and DFKI, Germany.

The appointed researchers are working on 15 individual, but related, projects which aim to improve the state of the art from five different directions: the user perspective, data collection and preparation, incorporation of language technology in translation memories, the human translator in the loop, and hybrid approaches to translation. To date, the project has produced over 120 high-quality publications and delivered three consortium wide training events. More details can be found in (Orasan et al. 2015).

References

Orasan, C.; Cattelan, A.; Corpas Pastor, G.; van Genabith, J.; Herranz, M.; Arevalillo, J; Liu, Q.; Sima'an, K.; Specia, L. (2015) The EXPERT Project: Advancing the State of the Art in Hybrid Translation Technologies. In Proceedings of Translating and the Computer 35, London, UK

Amplexor MTExpert - Machine Translation Adapted to the Translation Workflow

Alexandru CEAUSU, Sabine HUNSICKER, Tudy DROUMAGUET

AMPLEXOR International, Luxembourg, Luxembourg

Alexandru.Ceausu@amplexor.com, Sabine.Hunsicker@amplexor.com, Tudy.Droumaguet@amplexor.com

Abstract. MTExpert is AMPLEXOR's proprietary machine translation (MT) system based on state-of-the-art statistical and linguistic algorithms, easily integrated with existing linguistic assets, delivering quality results tailored to different communication objectives.

Description

AMPLEXOR MTExpert is a fully automated MT service based on the Moses open-source platform with *language-specific linguistic optimizations* (Ceausu and Hunsicker, 2014), as well as *terminology integration* and *format handling*. The available MT engines include specialised MT domains, as well as customisable MT engines adapted to a particular content type.

MTExpert provides translation for *specialised and generic MT language domains* for most European languages, Chinese, Japanese and Arabic. It includes domains ranging from the EU official publication domain to the technical domain or life sciences. Existing translation engines can be *customised* to a particular content type, based on existing language assets like translation memories or terminology databases.

MTExpert *confidence score* can automatically estimate the reliability of the machine translated content, for a better quality and cost-benefit assessment. The confidence score can be fine-tuned for each translation workflow using the feedback from translators (Hunsicker and Ceausu, 2015).

System integration into standard translation / post-editing processes is achieved using flexible interfaces such as CAT system plugins that allow a flexible threshold definition for the application of MT, e.g. for words in segments below a defined match quality. The output of MT is used as additional linguistic resource for translators and post-editors who can work in their usual translation environment.

References

Hunsicker, S., Ceausu, A. (2015) Machine Translation Quality Estimation Adapted to the Translation Workflow. Translating and the Computer 36: 133-136

Ceausu, A., Hunsicker, S. (2014) Pre-ordering of phrase-based machine translation input in translation workflow. LREC 2014: 3589-3592

Abu-MaTran: Automatic building of Machine Translation

Antonio TORAL[1], Sergio ORTIZ_ROJAS[2], Mikel FORCADA[3], Nikola LJUBESIC[4], Prokopis PROKOPIDIS[5]

[1]ADAPT Centre, School of Computing, Dublin City University, Ireland
[2]Prompsit Language Engineering SL, Spain
[3]Departament de Llenguatges i Sistemes Informatics, Universitat d'Alacant, Spain
[4]Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
[5]Athena Research and Innovation Center, Greece

atoral@computing.dcu.ie

Abstract. We present the current status of Abu-MaTran (http://www.abumatran.eu), a 4-year project (January 2013–December 2016) on rapid development of machine translation for under-resourced languages. It is funded under Marie Curie's Industry-Academia Partnerships and Pathways 2012 programme. This is a consortium-based project with 5 partners (4 academic and 1 industrial).

Description

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 (MT) by identifying crucial cutting-edge research techniques, 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: MT for the language of a new member state (Croatian) and related languages. All the resources produced are released as free/open-source software, resulting in effective knowledge transfer beyond the consortium.

At EAMT 2016 we will present a selection of the latests developments of the project: (i) state-of-the-art statistical and rule-based MT systems for South-Slavic languages based on free/open-source software, web crawled and publicly available data and linguistic knowledge, (ii) a novel tool for massive crawling of parallel and monolingual data from the Internet's top level domains and (iii) outcomes of the project's transfer and dissemination activities, e.g. MT hybridisation and web crawling for industry uses, rapid data creation for rule-based MT systems and establishment of a national linguistics Olympiad. All the resources developed within the project are freely available[1] and the MT systems deployed can be tested online.[2]

---

[1] http://www.abumatran.eu/?page_id=351

[2] http://translator.abumatran.eu/

TraMOOC (Translation for Massive Open Online Courses): Providing Reliable MT for MOOCs

Valia KORDONI, Lexi BIRCH, Ioana BULIGA, Kostadin CHOLAKOV, Markus EGG, Federico GASPARI, Yota GEORGAKOPOULOU, Maria GIALAMA, Iris HENDRICKX, Mitja JERMOL, Katia KERMANIDIS, Joss MOORKENS, Davor ORLIC, Michael PAPADOPOULOS, Maja POPOVIĆ, Rico SENNRICH, Vilelmini SOSONI, Dimitrios TSOUMAKOS, Antal van den BOSCH, Menno van ZAANEN, Andy WAY

Humboldt Universität zu Berlin (Coordinator, Germany), Dublin City University (Ireland), University of Edinburgh (UK), Ionian University (Greece), Stichting Katholieke Universiteit (Radboud University, Netherlands), EASN-Technology Innovation Services BVBA (Belgium), Deluxe Media Europe Ltd (UK), Stichting Katholieke Universiteit Brabant (Tilburg University, Netherlands), Iversity GMBH (Germany), Knowledge 4 All Foundation LBG (UK)

info@tramooc.eu

Abstract. TraMOOC is a 3-year EU-funded Horizon 2020 collaborative project that started in February 2015 (ICT-17-2014: Cracking the language barrier; project reference: 644333) which aims at providing reliable MT for MOOCs. The main outcome of the project will be a high-quality semi-automated MT service for all types of educational textual data available on a MOOC platform. The service will support nine European languages, in addition to Russian, and simplified Chinese. Progress may be monitored via the official project website at TraMOOC.eu, which is updated regularly with publications, news, and public deliverables.

TraMOOC: Reliable MT for Massive Open Online Courses

Language barriers constitute major impediments to sharing massive open online courses (MOOCs) with all peoples and in the attempt to educate all citizens. The EU-funded TraMOOC (Translation for Massive Open Online Courses) project aims at tackling this major issue by developing high-quality MT for all types of text included in MOOCs (e.g. assignments, tests, presentations, lecture subtitles, blog text) from English into eleven languages (DE, IT, PT, EL, DU, CS, BG, HR, PL, RU, ZH). The core of the TraMOOC platform will be open-source with some premium add-on services, enabling the integration of any MT solution in the educational domain, for any language. The project is conducting a major crowdsourcing drive, creating more training and evaluation data, while also performing comparative evaluations of phrase- and syntax-based SMT systems and state-of-the-art neural network MT engines, using standard metrics of fluency and adequacy, along with more detailed error taxonomies.

Modern MT: A New Open-Source Machine Translation Platform for the Translation Industry

U. GERMANN[1], E. BARBU[2], L. BENTIVOGLI[3], N. BERTOLDI[3], N. BOGOYCHEV[1], C. BUCK[1], D. CAROSELLI[2], L. CARVALHO[4], A. CATTELAN[2], R. CATTONI[3], M. CETTOLO[3], M. FEDERICO[3], B. HADDOW[1], D. MADL[1], L. MASTROSTEFANO[2], P. MATHUR[3], A. RUOPP[4], A. SAMIOTOU[4], V. SUDHARSHAN[4], M. TROMBETTI[2], J. van der MEER[4]

[1] University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom
[2] Translated srl, Via Nepal, 29, 00144 Rome, Italy
[3] Fondazione Bruno Kessler, Via Sommarive, 18, 38123 Povo, Italy
[4] TAUS B.V., Oosteinde 9, 1483 AB De Rijp, Netherlands

ugermann@inf.ed.ac.uk

Abstract. *Modern MT* (www.modernmt.eu) is a three-year Horizon 2020 *innovation action* (2015–2017) to develop new open-source machine translation technology for use in translation production environments, both fully automatic and as a back-end in interactive post-editing scenarios. Led by Translated srl, the project consortium also includes the Fondazione Bruno Kessler (FBK), the University of Edinburgh, and TAUS B.V. *Modern MT* has received funding from the *European Union's Horizon 2020 research and innovation programme* under Grant Agreement No645487 (call ICT-17-2014).

Project Description

Modern MT aims to improve the state of the art in open source machine translation software by developing cloud-ready software that offers

– A **simple installation** procedure for a ready-to-go, REST-based translation service.
– **Very fast set-up times** for systems built from scratch using existing parallel corpora (e.g., translation memories). The goal is to process incoming data at approximately the speed at which it is uploaded.
– **Immediate integration of new data** (e.g., from newly post-edited MT output). Rebuilding or retuning the system will not be necessary.
– **Instant domain adaptation** by considering translation context beyond the individual sentence, without the need for domain-specific custom engines.
– **High scalability** with respect to throughput, concurrent users, and the amount of data the system can handle.

A first version of the software is available at https://github.com/ModernMT/MMT.

Modern MT is also actively **collecting and curating parallel data** for internal use and public release from web crawls and contributions from translation stakeholders, to improve MT quality for everyone.

MODERN: Modeling Discourse Entities and Relations for Coherent Machine Translation*

A. POPESCU-BELIS[1], J. EVERS-VERMEUL[4], M. FISHEL[3], C. GRISOT[2], M.GROEN[4], J. HOEK[4], S. LOAICIGA[2], N.Q. LUONG[1], L. MASCARELL[3], T. MEYER[1], L. MICULICICH[1], J. MOESCHLER[2], X. PU[1], A. RIOS[3], T. SANDERS[4], M.VOLK[3], S. ZUFFEREY[4]

[1] Idiap Research Institute, 1920 Martigny, Switzerland
[2] University of Geneva, Department of Linguistics, 1211 Genève 4, Switzerland
[3] University of Zürich, Institute of Computational Linguistics, 8050 Zürich, Switzerland
[4] Utrecht University, Utrecht Institute of Linguistics, 3512 JK Utrecht, The Netherlands

andrei.popescu-belis@idiap.ch

Abstract. The MODERN project addresses coherence issues in sentence-by-sentence statistical MT, by propagating across sentences discourse-level information regarding discourse connectives, verb tenses, noun phrases and pronouns.

The goal of the MODERN project is to model and detect word dependencies across sentences, and to study their use by MT systems (MT), in order to demonstrate improvements in translation quality over state-of-the-art statistical MT, which still operates on a sentence-by-sentence basis. Three types of text-level dependencies are studied: referring expressions such as noun phrases and pronouns [4], discourse relations signaled by discourse connectives or implicit ones [1, 3], and verb tenses [2].

The overall approach of MODERN is to study the discourse-level phenomena from a theoretical perspective, but also using corpus-based approaches, in order to derive acceptable labels, features for automatic labeling, and training/test data. The automatic labeling systems are designed and combined with phrase-based statistical MT systems using factored models. The improvement in translating the respective phenomena is evaluated using specific automatic metrics or human evaluators.

MODERN started in 2013, building upon the COMTIS project started in 2010, with studies on English, French, German, Italian, Dutch, Arabic and Chinese. The main corpora used are Europarl, WIT3 (transcripts of TED talks), and Text+Berg (Swiss Alpine Club yearbooks).

References

1. Hoek J. et al., "The role of expectedness in the implicitation and explicitation of discourse relations", *Proc. of the 2nd DiscoMT workshop*, Lisbon, 2015.
2. Grisot C., *Temporal reference: empirical and theoretical perspectives*, PhD, UniGe, 2015.
3. Meyer T., *Discourse-level features for statistical machine translation*, PhD, EPFL, 2015.
4. Pu X., Mascarell L. et al., "Leveraging Compounds to Improve Noun Phrase Translation from Chinese and German", *Proc. of the ACL Student Session*, Beijing, 2015.

* MODERN is supported by the Swiss NSF, see www.idiap.ch/project/modern/.

Digital Curation Technologies (DKT)

Georg REHM, Felix SASAKI

DFKI GmbH, Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany

georg.rehm@dfki.de, felix.sasaki@dfki.de

Abstract. Digital Curation Technologies ("Digitale Kuratierungstechnologien", DKT) is a project that involves four Berlin-based SMEs (ART+COM AG, Condat AG, 3pc GmbH and Kreuzwerker GmbH) and DFKI GmbH (Language Technology Lab). The two-year action started in Sept. 2015 and aims at supporting digital curation processes, carried out by knowledge workers, through robust, precise and modular language and knowledge technologies. We combine these into workflows for the efficient processing, creation and dissemination of digital content. DFKI contributes language and knowledge technology components, including MT, and develops them further. Together with our partners, DFKI develops a platform for digital curation technologies, which offers services such as, e.g., translation, search, analytics, re-combination and summarisation. DKT is supported by the German Federal Ministry of Education and Research, Wachstumskern-Potenzial (no. 03WKP45). Details: http://www.digitale-kuratierung.de.

Description

The curation of digital content is a complex, knowledge- and time-intensive process in which authors or editors work, analyse, process and integrate highly heterogeneous data, content and media modules from disparate sources into a new content product that has a specific focus and communicative purpose. Among the processes involved are the selection, summarisation, classification, internationalisation, analysis, reordering, restructuring, and visualisation of various types of content. At the same time we need to bear in mind the continuously growing speed at which new information is coming in, the quantity and number of sources as well as of information pieces to be processed. The DKT SME partners contribute their respective sector-specific expertise (e.g., interactive showrooms, museums, television stations, newsrooms). In DKT we build showcases that demonstrate how language and knowledge technologies can be used in these typical digital curation workflows. The technologies contributed to DKT by DFKI belong to three areas: *semantic analysis* (e.g., information extraction), *semantic generation* (e.g., semantic story telling) and *multilingual technologies*, i.e., robust, adaptable MT components and integration of different monolingual or multilingual data and knowledge sources (including linked open data and standardised workflow data) for the sector-specific curation workflows. MT is used both for inbound and outbound translation (i.e., documents to be published). We also plan to incorporate MT engine improvement from user feedback. For the platform, we focus in particular the following features: fully integrated, robust, precise and scalable components with open interfaces in order to be efficiently embedded in sector-specific curation workflows; easy access to the cloud platform; application-oriented sector-technologies with high usability.

CRACKER – Cracking the Language Barrier. Selected Results 2015/2016

Georg REHM

DFKI GmbH, Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany

georg.rehm@dfki.de

Abstract. CRACKER is a Coordination and Support Action (ref. 645357; 01/2015–12/2017), funded by the EU. The consortium consists of seven partners: DFKI GmbH (DE); Charles Univ. in Prague (CZ); ELDA (FR); FBK (IT); R.C. "Athena" (GR); Univ. of Edinburgh (UK); Univ. of Sheffield (UK). Details are available at http://www.cracker-project.eu.

Description – Summary and Selected Results 2015/2016

The European MT research community is experiencing increased pressure for rapid success – from the political frameworks of the EU, but also from the business world. CRACKER pushes towards an improvement of MT research by implementing the successful example of other disciplines where massively collaborative research on shared resources – guided by interoperability, standardisation, challenges and success metrics – has led to important breakthroughs. The nucleus of this new R&D&I strategy is the group of projects funded through the call H2020-ICT-2014 topic 17, that is supported by CRACKER in coordination, evaluation and resources as well as in community building activities. CRACKER builds upon, consolidates and extends initiatives for collaborative MT research supported by earlier EU-projects.

In terms of selected results (2015/2016), a large number of evaluation tasks are now jointly organised by the IWSLT and WMT workshops. WMT went from five tasks in 2015 to ten tasks in 2016. IWSLT featured automatic speech recognition tasks (English, German), and spoken language translation and MT tasks involving English, German, French, Chinese, Czech, Thai, and Vietnamese. After a successful MT Marathon 2015 the event will again take place in Prague in September 2016, covering lectures, labs, projects, and tools. CRACKER co-organises, with the project QT21, an LREC 2016 Workshop on Translation Evaluation, bringing together representatives from language service providers and researchers working on high-quality MT, SMT and human evaluation. Resource sharing in CRACKER builds upon and extends the open resource exchange infrastructure META-SHARE. Its new version restructures the relevant content and aggregates it into one place; it also includes improved search as well as an updated licensing module. In terms of community building and outreach, CRACKER organised META-FORUM 2015 (and co-organised the Riga Summit 2015) and prepared, with the project LT_Observatory, a Strategic Agenda – an updated version will be presented at META-FORUM 2016 (Lisbon, July 4/5). CRACKER is also the main driver behind the emerging "Cracking the Language Barrier" federation of European organisations and projects working on technologies for multilingual Europe.