AI FOR TRANSLATION TECHNOLOGIES AND MULTILINGUAL EUROPE

Georg REHM

Joint work with Markus Foti and Josef Van Genabith

DFKI, Language Technology Lab, Saarbrücken, Germany

ABSTRACT

Europe is a multilingual society. In addition to the 24 official EU Member State languages, there are dozens of regional and minority languages as well as the languages of immigrants and important trade partners. This article briefly describes the multilingual setup of our continent, touching in particular on the grand challenges, such as establishing a Digital Single Market and the meaning and relevance Language Technologies can have for this flagship goal of the European Commission. Sophisticated Language Technologies, Machine Translation and other language-centric AI technologies can help radically transform our continent into a society in which multilingualism is fully enabled through ubiquitous digital technologies. The article provides a short description of the research results of the EU-funded research project QT21 in the area of Neural Machine Translation. It also highlights the European Commission's Connecting Europe Facility (CEF) programme, and especially its "Automated Translation" Digital Service Infrastructure (DSI), which will allow other DSIs and digital public services to be multilingual. The relevant activities are not only supported by the EC with regard

to public services but also steps towards supporting the Multilingual Digital Single Market. Furthermore, we briefly sketch the Cracking the Language Barrier federation as a European initiative that assembles many European organisations and projects working on technologies for multilingual Europe, as well as recent developments in the European Parliament. In the short to medium term, these may lead to establishing the Human Language Project as a pan-European funding programme for the next generation of Language Technologies and language-centric AI.

Introduction

In varietate concordia ("United in Diversity"), the EU motto, was officially adopted in 2000. It constitutes a rich tapestry of our diverse cultural, social, historical and linguistic experiences. The EU is committed to supporting this diversity. At the same time, and carefully balancing with the commitment to supporting diversity, the EU is striving towards overcoming barriers holding back economic and social development, caused by fragmentation into compliance, regulatory, legislative, business and linguistic silos, in order to support increased cross-EU exchange, mobility and trade to create opportunities, jobs, growth and wealth.

Languages are a key part of the European identity and diversity (EU Charter, Art. 22; Treaty on European Union, Art. 3). The EU currently has 24 official languages that all share the same status; multilingualism is at the very heart of the European idea. In addition to the 24 official ones, there are dozens of regional and minority languages as well as languages of immigrants and trade partners. Yet while languages are key parts of our rich cultural identities, they can also constitute barriers to the free flow of people, ideas, commerce, trade, administration and cultural exchanges across linguistic boundaries. For example, in terms of economic challenges, if the Digital Single Market is not inherently multilingual, the EU will end up with more than 20 isolated digital markets, strictly separated by their linguistic boundaries. In this respect, language barriers are market barriers. In terms of social and public challenges, all EU citizens (and beyond) should be empowered to use their mother tongues online and in all other forms of cross-border, cross-lingual and cross-cultural digital communication, including regional, national and international digital public services. Most of us are not usually able to master more than two or three languages. Language Technologies (LTs) can help overcome language silos, especially in combination with modern AI technologies .

In 2010, recognising Europe's demand and opportunities, 60 research centres in 34 European countries joined forces in META-NET, a European Network of Excellence dedicated to the technological foundations of a multilingual, inclusive and innovative

European society.[1] META-NET was partially supported by four EU projects (T4ME, CESAR, METANET4U, META-NORD, 2010–2013); until recently, CRACKER (2015-2017) supported selected META-NET activities such as, among others, META-FORUM 2015, 2016 and 2017 as well as META-SHARE maintenance.[2] One of META-NET's key goals is technology support for all European languages as well as fostering innovative research by providing strategic guidance and recommendations with regard to key research topics. Crucial milestones were the publication of the META-NET White Papers (Rehm and Uszkoreit, 2012; Rehm et al., 2014) and the Strategic Research Agenda for Multilingual Europe 2020 (SRA) (Rehm and Uszkoreit, 2013), the implementation and deployment of META-SHARE as well as the publication of three versions of the Strategic Research and Innovation Agenda for the Multilingual Digital Single Market (Rehm, 2015; Rehm, 2016b, Rehm 2017). The impact of these activities in various European countries is documented in (Rehm et al., 2016a; Rehm et al., 2016b). At the same time, Language Technologies and Artificial Intelligence in general have been making unprecedented progress, especially regarding technologies based on Neural Networks (NNs). The success of Deep Learning technologies is based on the availability of very large data sets coupled with the sheer computing power and speed of modern IT hardware and powerful neural algorithms. In 2016/2017, AI-based technologies have started penetrating every single sector, leading to large social and economic disruptions, from self-driving cars to different types of robots, from image recognition and manipulation to machine translation.

In the remainder of this article, we first outline key results produced by the EU-funded MT research project QT21 (Section 2). In Section 3 we concentrate on the Connecting Europe Facility (CEF) programme, especially the eTranslation activity that helps to ensure that Europe's digital public services will be multilingual. Section 4 describes the Cracking the Language Barrier federation that is forging the Multilingual Europe community. In Section 5 we briefly touch upon a workshop on LT held in the European Parliament in early 2017. Section 6 provides a brief overview of the vision of the Human Language Project, while Section 7 concludes the article.

1. European MT Research – Results from QT21

Morphologically rich and syntactically varied languages have, for a long time, presented particular challenges to MT. A single word may appear in many different forms (inflections), expressing grammatical properties including case, number, gender, tense, aspect, modality etc.; some words that are related in a sentence may have to agree with respect to some of their morphological properties

[1] http://www.meta-net.eu

[2] http://www.cracker-project.eu

(congruence), and it may be possible to articulate the same proposition in many different ways (word order) using the same words. Modern MT is generally data-driven and machine-learning based. For morphologically rich and syntactically varied languages, even large sets of training data may not capture the full variability in terms of diversity of word forms and syntactic possibilities the language is capable of. This can have adverse effects on the quality of the MT output, especially when translating into morphologically rich languages. Indeed, MT quality for translating, e.g., from English into German, Czech, or Hungarian tends to be lower compared to the quality from English into French or Spanish, given the same amount of training data (number of parallel sentences).

The project QT21[3] targeted exactly those challenging linguistic and resource scenarios: morphologically rich and syntactically varied languages, focusing on key representatives of the major European language families (Romance, Germanic and Slavic: Romanian, German and Czech) as well as a representative of a smaller European language family (Baltic: Latvian), and English, as a morphologically and syntactically more restricted language. For evaluation, the project focused on international shared tasks, competitions where MT teams from across the globe compare their systems in evaluation campaigns. The most important such competition is WMT.[4] The QT21 team assembled some of the top European research teams in MT (University of Edinburgh, RWTH Aachen, KIT Karlsruhe, University of Amsterdam, University of Sheffield, LIMSI, Dublin City University, Charles University Prague, FBK and DFKI) as well as industry partners active in research (Tilde, text&form and TAUS).

When QT21 started, statistical approaches to machine translation (SMT) defined the state-of-the-art in MT. SMT learns the translation of individual words and word chunks (called "phrases", consisting of two or more consecutive words, and generally up to a maximum of seven words) from bilingual data and estimates the probability of such translation blocks. Given a new sentence to be translated, SMT follows a divide-and-conquer strategy: it chunks the input sentence into smaller bits and pieces (phrases, including individual words), finds the translations of them in its database together with their associated probabilities, and pieces the translation pieces together to form an overall translation. This may include some reordering (the order of the corresponding words and phrases in source and target may differ) and a target language model (LM). The target LM is a large list of word sequences found in very large target language data corpora accompanied by probabilities of the words that, given such a sequence, can follow them. This can help distinguish grammatical from ungrammatical sequences. For example, the probability that the word "computer" follows the word "a" should be higher than that of the plural word, "computers": $P(computer|a) > P(computers|a)$. An SMT system

[3] http://qt21.eu

[4] http://www.statmt.org/wmt18/

uses translation and language model probabilities (as well as other probabilities such as reordering probabilities) to find the best translation for a given input sentence under the model. While SMT models have been highly successful, in coming up with a translation they make many local decisions and put these together to form the overall translation. Mathematically speaking, they make many statistical independence assumptions in carrying out these local decisions. This is not always optimal, as decisions that may seem reasonably local, may not always make the best contribution towards finding the overall best solution. What is more, the components of an SMT system, including the translation, reordering and language models, are often estimated differently and independently of each other and then put together into the overall system using weights, which are optimised in a final training step. Again, this may not be optimal. Nevertheless, SMT systems have been well honed and engineered over more than three decades (Brown et al. 1988; Koehn et al. 2003; Chiang 2005; Koehn 2010; Williams et al. 2016) and with many important and clever extensions and improvements were the dominant MT approach, as recent as in the international MT competitions in 2015 (WMT 2015).

Compared with SMT, Neural MT (NMT) is the new kid on the block. Apart from early work (Forcada and Ñeco 1997), NMT publications started emerging around 2012-2015 (Schwenk 2012; Kalchbrenner and Blunsom 2013; Cho et al. 2014a; Cho et al. 2014b; Sutskever et al. 2014; Bahdanau et al. 2015). By 2015, NMT was still generally out-performed by SMT (in fact by the QT21 SMT systems, WMT 2015). However, the QT21 team quickly recognised the potential of NMT, and developed important improvements to NMT, including byte-pair-based sub-word representations (Sennrich et al. 2016a) (particularly important for morphologically rich languages) and back-translation (Sennrich et al. 2016b) (particularly important for challenging resource scenarios), that made NMT comprehensively outperform SMT approaches at WMT 2016 (again, as in 2015, the majority of the shared tasks was won by QT21 engines, in particular from Edinburgh, but this time with NMT engines). QT21 repeated the feat, again winning the majority of the shared tasks in 2017, again with NMT engines. The combined impact from the shared tasks was such that SMT technology has by now almost completely been replaced by NMT approaches.

Compared with SMT, NMT is more "holistic": an encoder-decoder based NMT system that uses recurrent NNs, e.g., "reads in" the full input sentence and maps it into a neural representation of the sentence (a dense vector in some fairly low dimensional vector space – with a few hundred dimensions), and then passes on that representation to a decoder, that "spells out" that representation word by word into a sentence in the target language. The resulting target sentence is hopefully a good translation of the input sentence, as the complete encoder-decoder system has been trained jointly end-to-end using source-target sentence pairs that are translations of each other.

This optimises the internal representation "handed-over" from encoder to decoder to capture what is important for producing good translations. Decisions in this type of NMT happen against the global context of what needs to be done: a translation-relevant representation of the input sentence, rather than being decomposed into many small local and independent decisions as in SMT. This is one of the reasons why NMT translations are often perceived to be more fluent than SMT translations.

QT21 has made key contributions to the paradigm shift away from SMT to NMT. In addition, QT21 made important contributions to multi-modal NMT, where MT systems use information from images and source text in translation (Calixto et al. 2017); multilingual NMT, where systems learn to translate between many languages simultaneously, allowing information to be shared across languages, with benefits to in particular low resource languages (Ha et al. 2016); integrating knowledge graph information into NMT and multi-lingual NMT, enabling "beyond-zero-shot translation" (España-Bonet and van Genabith 2017); noise-robust NMT (Heigold et al. 2018); automatic post-editing (Chatterjee et al 2017a, Chatterjee et al 2017b) and continuous learning from corrections (Turchi et al. 2017); and MT evaluation and metrics, including the Direct Assessment (Graham and Liu 2016) method, improved metrics and new test suites (Macketanz et al. 2018).

2. Connecting Europe Facility – Automated Translation

2.1. The European Commission and Machine Translation

The European Commission, along with the other EU Institutions, has been very active in the machine translation field, which is only to be expected of an organisation where multilingualism plays such a crucial part. When it comes to European policy making, translation needs to be both accurate and quick, so the Directorate-General for Translation (DGT) carefully observes trends in the technology, as it does for all translation aids.

DGT stores most of its translations within a central translation memory to which all translators can refer. This database, called Euramis, is in itself a fantastic resource, but in addition, the existence of such a large and constantly updated trove of multilingual data means that DGT has at its disposal a huge multilingual corpus that is ideal for the big data revolution and MT.

Accordingly, in 2010 the idea was born to use this data to build an MT system with the statistical approach which was then state-of-the-art, using the MOSES toolkit developed in part thanks to EC research funding. The fully-fledged system, covering all EU languages, was unveiled on 3 July 2013. But a computer and a piece of software are not enough to produce a proper MT system. Expertise is also needed, and not only the technical skills to put the pieces together and build a system that can spit out translations.

DGT is in a near-unique position of having a vast trove of translation expertise and stylistic decisions honed over years, along with a body of highly-skilled professionals with the attention to detail needed to spot issues. The MT@EC technical team worked closely with the different language departments to determine standard expressions and styles for dates, references, even punctuation – apparently in Denmark the question of which quotation marks are the proper ones remains unresolved, and some English speakers have strong opinions about whether apolog*ise* or apolog*ize* is the correct spelling, to say nothing of spelling reforms for Maltese, Portuguese and German. In all of these cases, DGT called upon the expertise of native speakers to make the correct choices and adapt MT@EC to current usage.

Though developed in house at DGT, the system was intended to be used by all of the EU Institutions, as well as by public administrations in the Member States. It would – and has – helped to cope with an increased translation workload while staff numbers were being reduced. But the availability of a secure and reliable MT system within the EU Institutions has also given EU staff and public administrations a tool to get the gist of documents that can, in some cases, be hundreds of pages long, and either get the information they need without requesting a human translator to deal with a document over weeks, or else better target the need for translation by identifying sections that are relevant to the issue at hand and requesting accurate, verified human translation only for the 10 or 20 pages actually needed.

Moreover, despite the fact that English's role as a lingua franca continues to grow in the digital age, the need and desire for translation is also growing. Despite the European Union's commitment to multilingualism, many of its own websites have not been able to provide versions in all EU languages, partly because the human translators available are working on more critical legal or political documents, and partly because the content of modern websites is dynamically changing day-by-day, hour-by-hour, and even minute-by-minute on social media which enables the public to give feedback directly.

To meet this need, MT@EC therefore also offered translation via a web service so that websites could translate content, albeit with the caveat that it should always be identified as machine translation and be requested by the user. There is nothing worse than reading a website where the stilted language makes it slowly dawn on you that your language doesn't deserve a full, human quality translation, unless perhaps it is being served up an incomprehensible word salad that claims to be Finnish or Hungarian!

With the mobile revolution, MT has become a common tool, and with the continuing advances in computing power and the rise of big data, the statistical system that had served the EU Institutions well for several years risked becoming obsolete. Fortunately, the Juncker Commission had named the Digital Single Market as one of its priorities, and along with the desire to foster innovation, create faster Internet connections and cross-border digital interaction, came a need for language technologies. Machine translation and the MT@EC system was identified as a key Digital Building Block, along with software such as eID, eDelivery and eInvoicing, to ensure trustworthy network computing across the EU.

In order to achieve these goals, from 2015, this digital component of the Connecting Europe Facility (CEF) was set up as a broad programme with the Directorate-General for Communications Networks, Content and Technology (CNECT) in charge of the project as a whole, and DGT providing its expertise and experience to upgrading the MT system into something capable of supporting the broader ambitions of CEF. Thus was born CEF.AT (*automated translation*) and eTranslation as the translation system itself.

The CEF programme provided funding to upgrade both the translation technology – in the short time since MT@EC was launched, statistical machine translation was beginning to lose ground to the new kid on the block – neural machine translation. One-by-one, the big players were switching to this new technology, and the infrastructure to run it.

An expanded team at DGT led to the launch of MT@EC's successor, eTranslation, on 3 July 2017 as a limited system for plain text and machine-to-machine use only – but already the Commission's first neural engine with English into Hungarian. The full system followed on 15 November, incorporating document processing for common formats (Microsoft, Libre Office, PDF) and especially, the first wave of neural engines: English into and out of German, Estonian, Finnish and Hungarian. The migration to cutting edge technology had begun! The languages which were most poorly served by statistical technology were the first to benefit from the new technology, and will be followed through 2018 by more Baltic and Slavic languages, before the Germanic and Latin families are tackled.

But the CEF.AT platform is not only an MT system. EU translations are undoubtedly of high quality, but they have a certain legalistic flavour which feeds through to eTranslation, making it a fantastic tool for EU documents, but sometimes stilted for newspaper articles or common speech. And yet the Digital Single Market is striving to get people to, for example, make purchases on websites based in other countries, secure in the knowledge that if there is a problem, they will be able to use services such as the Online Dispute Resolution system to resolve them in a language they are less comfortable in.

For eTranslation to help in such cases, it needs a broader range of data to produce translations more suited to that style.
For this reason, the European Language Resource Collection (ELRC) service contract was set up and tasked with running

workshops in all EU countries to build up a network of data sources to make eTranslation more useful to the European public, as well as to raise awareness of the system and boost usage. These workshops were held through 2015 and early 2016, with a second round launched in 2017 and 2018. The data gathered will be used by eTranslation, but it will also be a product in and of itself, available to the general public for research purposes.

2.2. Data for CEF eTranslation: European Language Resource Coordination

The Digital Single Market (DSM) is one of the key priorities of the EU 2014-2020. It is estimated that a successfully and fully implemented DSM could contribute €415 billion p.a. to the EU economy. The DSM is supported by several actions, including the CEF programme.[5] CEF is designed to help overcome fragmentation and the resulting barriers for citizens, businesses and public services to fully participate in cross-EU social, economic and cultural opportunities and exchange. CEF focuses on Energy, Telecom and Transport, its funding totals €30.4 billion for 2014-2020. CEF is a "funding instrument to promote growth, jobs and competitiveness through targeted infrastructure investment at European level", it "facilitates cross-border interaction between public administrations, businesses and citizens" (ibid.).

CEF is implemented in terms of Digital Service Infrastructures (DSIs) and Building Blocks (BBs). DSIs are sector specific and include services targeting (mainly, but not exclusively): Justice, Home Affairs and Citizens' Rights (eJustice; Online Dispute Resolution, ODR; Safer Internet and Cybersecurity), Science (Public Open Data), Business (Business Registers Interconnection System, BRIS; eProcurement), Employment and Social Rights (Electronic Exchange of Social Service Information, EESSI), Health (eHealth) and Culture (Europeana).

CEF Building Blocks are general CEF DSI services that are re-used in other DSIs. These Building Blocks include eID, eSignature, eInvoicing, eDelivery and eTranslation.

CEF eTranslation[6] is a corner stone building block DSI based on MT designed to support multilingual access and interaction with CEF DSIs. Importantly, CEF goes beyond the official EU languages and includes the languages of the CEF-affiliated countries Norway and Iceland.

A number of service contracts (under CEF SMART 2014, 2015 and 2016) were

[5] https://ec.europa.eu/digital-single-market/en/connecting-europe-facility

[6] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

initiated to support the collection of relevant training data (ELRC) for CEF across Europe, explore LT technology provision for procurement under CEF, to gauge the structure and size of the LT market in Europe etc., as well as integration and implementation projects, e.g., under CEF-TC-2016-3, including contracts on crawling corpora (Provision of Web-Scale Parallel Corpora for Official European Languages), terminology (eTranslation TermBank), domain-specific applications (CEF Automated Translation for the EU Council Presidency, Cross-border eProcurement notifications, etc.), resources for number of languages (European Language Resource Infrastructure, ELRI).

In the following, we describe the work carried out under the European Language Resource Coordination (ELRC) service contract (Lösch et al. 2018).[7] ELRC is coordinated by DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz, Germany), in partnership with ELDA (Evaluations and Language Resources Distribution Agency, France), ILSP/R.C. "Athena" (Institute for Language and Speech Processing, Greece) and Tilde (Latvia).

In order to support access to CEF DSIs across official EU and CEF-affiliated languages, eTranslation uses MT. Modern MT is data-driven and machine-learning based. To produce the best translations possible, the engines need the right kind of training data. While the EU has ample and good training data to translate European parliament debates (EuroParl) or EU Laws (JRC Acquis), currently it does not have the right kind of training data to support CEF DSIs in verticals as diverse as health, consumer rights and culture (see DSIs above). In order to acquire the right kind of training data, ELRC is targeting national public services in the EU Member States and CEF-affiliated countries.

ELRC is carrying out this work under the PSI Directive. In accordance with the Public Sector Information (PSI) Directive 2003/98/EC (modified in 2013 by the Directive 2013/37/UE), Member States should ensure that documents, which are held by public sector bodies and accessible according to national access regimes, are re-usable for commercial or non-commercial purposes. ELRC is seeking domain/vertical specific bi-lingual and mono-lingual data (including terminologies) for training MT systems for CEF.

The work carried out under ELRC is based on the following guiding and strategic principles: local ownership, awareness raising, building an infrastructure for data collection, carrying out the data collection, making data collection sustainable, showing and demonstrating the impact and reward of the data. These principles are further discussed below.

Local Ownership: ELRC is about data sharing and mutual benefits from sharing data. The starting point is to establish local ownership of the data sharing process in each EU and CEF-affiliated country ("Supporting the local language is

[7] http://lr-coordination.eu

supporting the EU and supporting the EU is supporting the local language"). To achieve this in each country we have established a pair of National Anchor Points (NAPs), one technology NAP (TNAP) who is a leading local expert in language technologies and has excellent reach into national public services and political decision making processes, and a public service NAP (PNAP) who is a leading member of the national public services with a strong interest in languages and language technologies and is exceptionally well connected with language and language technology supporters in other similar national public services. Together the NAPs form the European Language Resource Board (LRB), to serve as governance and oversight body for the ELRC effort.

ELRC initially focused on awareness raising: many national public services hold translation data resulting from previous translation activities. However, many national public services are not fully aware of the value of the data they hold.

Building an infrastructure for data collection: the ELRC consortium built ELRC-SHARE (Piperidis et al. 2018), based on META-SHARE (Piperidis 2012), to host and clear data donations. In addition to hosting the data, this includes basic pre-processing such as data conversion (e.g., to plain text or XML), tag removal, re-formatting, data extraction, cleaning and alignment, meta-data validation and anonymization as well as IP clearance to clarify legal issues linked to IPR and licensing constraints.

Carrying out the data collection: ELRC started collecting resources in the spring of 2016. Within the first year, ELRC managed to collect 225 Language Resources (LRs), covering all official EU languages, plus Icelandic and both varieties of Norwegian, Bokmål and Nynorsk. Data collection is on-going.

Making data collection sustainable: ELRC is currently working with the NAPs on strengthening and developing structures inside the EU membership and CEF-affiliated countries to make the data pipeline and the mutual benefit for the local language through better language technology support and the EU through increased language transparency for CEF sustainable beyond the lifetime of the ELRC service contract.

Showing the impact and reward of the data: ELRC works closely with colleagues at eTranslation at DGT to make MT engines trained with EU Member State and CEF-affiliated countries contributions available to donating and other public services in these states, to show the rewards and improvements obtained by the data.

Finally, and importantly, activities under eTranslate, ELRC and CEF SMART are in no way meant to distort commercial business opportunities. On the contrary, they are all without exception part of a long-term strategy by the EC to open up market places and business opportunities for growth, employment, and the creation of wealth in the EU in the crucial areas of language technologies and services: all public data collected by ELRC is already or going to be made available publicly for both commercial and research activities. SMART programmes are under way to scout the commercial EU language technology and services market for public contracts and procurement.

3. The Cracking the Language Barrier Federation

The *Cracking the Language Barrier* initiative is a federation of projects and organisations working on technologies for a multilingual Europe (Rehm, 2016a).[8] It was established as part of the activities of CRACKER, starting out with the group of projects funded through the call ICT-17-2014. The federation is meant to serve as an umbrella initiative that includes all currently running and recently completed EU- funded projects and, in particular, all stakeholder organisations. The term "federation" emphasises that this is an initiative *from* the community *for* the community. All members have equal rights and equal say.

The objective of "cracking the language barrier" (or working with or on multilingual or crosslingual technologies) is the shared strategic goal that all members firmly stand behind. The group of members is constantly growing. The initiative is meant to be a self-organising federation of projects and organisations that share a common strategic objective. It is currently not foreseen to establish a governance structure (Rehm et al., 2016a). At the time of writing, the federation has as its members 12 organisations and 25 projects. Areas of collaboration include external communication and dissemination, data management and repositories, tools and technologies, shared tasks and evaluations as well as SRIA development.

[8] http://www.cracking-the-language-barrier.eu

4. European Parliament STOA Workshop (January 2017)

On 10 January 2017 the workshop "Language equality in the Digital Age – Towards a Human Language Project" took place in the European Parliament in Brussels.[9] The workshop was informed and made possible through the work META-NET conducted in the White Paper Series and the alarming observation that at least 21 European languages are in danger of digital extinction (Rehm and Uszkoreit, 2012; Rehm et al., 2014). Several presentations provided insight into specific areas, barriers and related topics. The participants also discussed potential opportunities and solutions. The workshop was organised by the Science and Technologies Options Assessment panel (STOA).

The title of the event already hinted at a possible next step, which was also reflected and suggested in as well as supported by

[9] http://www.stoa.europarl.europa.eu/stoa/cms/home/workshops/language

several of the presentations, i.e., to set up, with the help of the European Commission, European Parliament and Member States, a large-scale, long-term flagship initiative to carry out research, development and innovation activities (including education and commercialisation) with substantial funding towards a common strategic goal. Under the umbrella of this Human Language Project (HLP) new breakthroughs are to be made in order to address the threat of digital language extinction but also to provide solutions to the European citizen, industry and administrations. The participants supported the idea, made in one of the presentations, to move forward with the highly ambitious goal of reaching *Deep Natural Language Understanding and Generation by 2030*. All participants of the workshop, including several members of the EP, supported the idea of setting up the HLP. After the workshop, in March 2017, a corresponding study was published (STOA, 2017), commissioned by the EP. It provides 11 policy recommendations, most importantly, to set up the HLP.

5. Towards the Human Language Project

As can be seen from the short summaries above, current developments are moving into the right direction with regard to the goal of establishing a technology base for multilingual Europe. The growth in AI is producing effective tailwind for our field, not only with regard to solutions for current challenges, such as robust interactive systems for connected devices, but also to open doors to the often neglected topic of Language Technologies. CEF AT is making progress and extended through additional use case projects. Not only in CEF but also in many of the EU Member States and other European countries there are projects working either on the foundations or on applied LT projects. Some European countries have already or are actively working on establishing national LT-related funding programmes, for example, Spain.

In addition to the above mentioned short to medium-term opportunities, we recommend to the whole European Computational Linguistics and Language Technology community to collaborate closely together in order to initiate the Human Language Project as a long-term, large-scale, sufficiently funded EU flagship initiative and unprecedented opportunity for Europe to work on the next generation of Language Technologies. As the key scientific goal it was recommended to strive for *Deep Natural Language Understanding and Generation by 2030*, especially by collaborating closely between Computational Linguistics, Linguistics, Artificial Intelligence, Machine Learning and Knowledge Technology. We foresee setting up a shared programme between the EU (crucially, through the framework programme that will succeed Horizon 2020) and the Member States and other stakeholders, especially industry. The setup needs to include an intertwined mix of basic research, applied research, technology development, innovation and commercialisation; education and talent retention also need to be taken into account. The HLP should run for at least ten years, ideally 15 years, so that the ambitious scientific goal can be adequately addressed. Public procurement and a language-related policy change towards "LT-enabled multilingualism" are crucial related aspects. Additional details can be found in the STOA study (STOA, 2017) and the most recent version of the Strategic Research and Innovation Agenda for European LT (Rehm, 2017).

6. Conclusions

For several years we have been witnessing an intense uptake of Language Technologies (LT) in our day-to-day IT infrastructure, from personal assistants in connected devices to the Internet of Things and mobile phones, among others. Many of the current innovations have their roots in European research activities, but most of them are currently deployed by SMEs or larger enterprises on other continents. Now that LT is receiving more and more attention and uptake globally, there is an urgent need for the wider LT community to act. There is also a need to think about what the European perspective is, if Europe can afford *not* to invest in this topic and also what Europe can gain from supporting Language Technologies. With the Human Language Project, the EU has a unique opportunity significantly to invest in a set of basic and applied technologies that will be nothing but crucial and mission-critical for the next generations of IT. While Europe has missed several of the recent IT trends, Language Technology is a unique opportunity for Europe to 'invent' and support its own much needed flagship initiative instead of following the priorities set out by other continents and countries. The Human Language Project could not only bring about novel breakthroughs and paradigm-shifting research results, it would also produce technologies that would benefit European society, administrations and companies. The European LT community recently published the final version of the Strategic Research and Innovation Agenda, "Language Technologies for Multilingual Europe: Towards a Human Language Project" (Rehm, 2017). In the document we included suggestions for establishing bridges to neighbouring scientific fields, for example, Artificial Intelligence, Cognitive Science, Cognitive Computing, Machine Learning and Deep Learning, Linked Data, Psychology and Digital Humanities and also to important IT topics such as, for example, Robotics, Internet of Things, Smart Manufacturing and Smart Cities.

With enough substance in terms of strategies, technology solutions and innovative research topics as well as with the collective weight and support of all involved communities and as many countries as possible, we may have a chance to help establish the Human Language Project (Rehm and Hegele, 2018). With regard to the Cracking the Language

Barrier federation and the increasing number of European and national projects, there is a need for an adjustment of the overall structure. Not only should national research projects and national organisations (for example, national Computational Linguistics associations) be taken on board as members of the federation, we should also make use of the federation to showcase European research results and technological solutions. In our communication and outreach activities, we need to work on closing the gap between AI and LT, making the point that LT is an inherent component of AI and vice versa. Additionally, our community should also benefit from the interest in AI by concentrating on language as the next goal to tackle with the help of AI-based approaches.

Since 2012 we have been arguing that the Digital Single Market (DSM) concept and strategy needs to take into account the multilingualism and language topic in order for the DSM to become a truly unified *single* market. By now the EC has acknowledged the importance in several communications, which is why additional activities with regard to establishing the technology layer of the Multilingual DSM can be expected in the future, maybe also a closer collaboration with the CEF AT programme.[10] CEF AT is critically important not only to enable multilingual digital public services but also to showcase current European LT research and deployment activities and their benefits. It needs to be further supported and tightly integrated into our strategic plans due to its importance for the task of awareness raising and informing key decision makers about the capabilities of "Language Technology made in Europe".

Technologies based on AI methods are currently penetrating and disrupting all industries and sectors – including the field of translation: from the very large translation departments of the European Institutions to freelancers and the thousands of SMEs that offer and carry out language services throughout Europe. While the different academic and commercial research teams working on Neural MT are continuously producing breakthrough after breakthrough, perfect automatic machine translation is still very far away. One important and relevant consequence is not only a constantly improving set of tools for gist translation, translators also benefit from a vastly improved technology and tool landscape that will provide them with a multitude of adaptive tools and smart dictionary and terminology services. The goal of the Human Language Project is to secure Europe's pole position in this field and to produce the next overall scientific breakthroughs that Computational Linguistics and Language Technology have been working on for decades: Deep Natural Language Understanding.

[10] See, for example, https://ec.europa.eu/commission/commissioners/2014-2019/ansip/blog/how-multilingual-europes-digital-single-market_en and https://ec.europa.eu/digital-single-market/en/blog/multilingualism-digital-age-barrier-or-opportunity

References

Bahdanau, D., K. Cho, and Y. Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: Proceedings of the International Conference on Learning Representations (ICLR)

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. L. Mercer, P. Roossin (1988). "A statistical approach to language translation". COLING' 88. Association for Computational Linguistics. 1: 71–76

Calixto, I., Q. Liu, N. Campbell: "Doubly-Attentive Decoder for Multi-modal Neural Machine Translation". Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pages 1913–1924, Vancouver, Canada, July 30 - August 4, 2017

Chatterjee, R., G. Gebremelak, M. Negri, M. Turchi (2017a): "Online automatic post-editing for MT in a multi-domain translation environment." In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (Vol. 1, pp. 525-535).

Chatterjee, R., M.A. Farajian, M. Negri, M. Turchi, A. Srivastava, S. Pal (2017b): "Multi-source Neural Automatic Post-Editing: FBK's participation in the WMT 2017 APE shared task". In Proceedings of the Second Conference on Machine Translation (pp. 630-638).

Chiang, D. (2005). "A Hierarchical Phrase-Based Model for Statistical Machine Translation". Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). Pp. 263—270

Cho, K., B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, Y. Bengio (2014a): "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, October 25-29, 2014, Doha, Qatar.

Cho, K., B. van Merrienboer, D. Bahdanau, Y. Bengio (2014b): "On the properties of neural machine translation: Encoder–Decoder approaches". In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 103–111, October 25, 2014, Doha, Qatar.

España-Bonet, C., J. van Genabith (2017): "Going beyond zero-shot MT: combining phonological, morphological and semantic factors". The UdS-DFKI System at IWSLT 2017. Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT), pages 15-22, Tokyo, Japan, December 2017.

Forcada M.L., R.P. Ñeco (1997): "Recursive hetero-associative memories for translation." In: Mira J., Moreno-Díaz R., Cabestany J. (eds) Biological and Artificial Computation: From Neuroscience to Technology. IWANN 1997. Lecture Notes in Computer Science, vol 1240. Springer, Berlin, Heidelberg

Graham, Y., Q. Liu (2016): "Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics." In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), San Diego, CA.

Ha, T.-L., J. Niehues, A. Waibel (2016): "Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder". Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT), Seattle, USA, December 8-9, 2016

Heigold, G., S. Varanasi, G. Neumann, J. van Genabith (2018): "How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse?" Association for Machine Translation in the Americas (AMTA) 2018, Boston, Massachusetts, USA, AMTA, 2018

Kalchbrenner, N., P. Blunsom (2013): "Recurrent continuous translation models". In Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1700–1709. Association for Computational Linguistics.

Koehn, P., F.J. Och, D. Marcu (2003): "Statistical Phrase-Based Translation". Proceedings of HLT-NAACL, Edmonton, pp. 48-54

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Lakew, S.M., Q.F. Lotito , M. Negri, M. Turchi, M. Federico (2017): "Improving Zero-Shot Translation of Low-Resource Languages", Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT), Tokyo, Japan, December 2017. (Best Student Paper Award)

Lösch, A., V. Mapelli, S. Piperidis, A. Vasiļjevs, L. Smal, T. Declerck, E. Schnur, K. Choukri, J. van Genabith (2018): "European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management". LREC 2018. Miyazaki. Japan.

Macketanz, V., R. Ai, A. Burchardt (2018): "TQ-AutoTest – An Automated Test Suite for (Machine) Translation Quality", LREC 2018, Miyazaki, Japan

Piperidis, S. (2012): "The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions". In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rehm, G., H. Uszkoreit (eds.) (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, Heidelberg, New York, Dordrecht, London. 31 volumes on 30 European languages. http://www.meta-net.eu/whitepapers.

Rehm, G., H. Uszkoreit (eds.) (2013). *The META-NET Strategic Research Agenda for Multilingual Europe*. Springer, Heidelberg, New York, Dordrecht, London. http://www.meta-net.eu/sra.

Rehm, G., H. Uszkoreit, I. Dagan, V. Goetcherian, M.U. Dogan, C. Mermer, T. Váradi, S. Kirchmeier-Andersen, G. Stickel, M.P. Jones, S. Oeter, S. Gramstad (2014): "An Update and Extension of the META-NET Study *Europe's Languages in the Digital Age*." In *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, Reykjavik, Iceland, May.

Rehm, G., J. Hajic, J. van Genabith, A. Vasiļjevs (2016a): "Fostering the Next Generation of European Language Technology: Recent Developments – Emerging Initiatives – Challenges and Opportunities". In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 1586–1592, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Rehm, G., H. Uszkoreit, S. Ananiadou, N. Bel, A. Bielevičienė, L. Borin, A. Branco, G. Budin, N. Calzolari, W. Daelemans, R. Garabík, M. Grobelnik, C. García-Mateo, J. van Genabith, J. Hajič, I. Hernáez, J. Judge, S. Koeva, S. Krek, C. Krstev, K. Lindén, B. Magnini, J. Mariani, J. McNaught, M. Melero, M. Monachini, A. Moreno, J. Odjik, M. Ogrodniczuk, P. Pęzik, S. Piperidis, A. Przepiórkowski, E. Rögnvaldsson, M. Rosner, B. Pedersen, I. Skadiņa, K. DeSmedt, M. Tadić, P. Thompson, D. Tufiş, T. Váradi, A. Vasiļjevs, K. Vider, J. Zabarskaite (2016b): "The Strategic Impact of META-NET on the Regional, National and International Level". *Language Resources and Evaluation*. 10.1007/s10579-015-9333- 4.

Rehm, G. (ed.) (2015). *Strategic Agenda for the Multilingual Digital Single Market – Technologies for Overcoming Language Barriers towards a truly integrated European Online Market. April. Version 0.5. April 22, 2015. Prepared by the EU-funded projects CRACKER and LT_Observatory.*

Rehm, G. (2016a): "Cracking the Language Barrier for a Multilingual Europe". In Gerhard Stickel et al., editors, *Language Use in Public Administration. Contributions to the Annual Conference 2015 of EFNIL in Helsinki*. Peter Lang, Frankfurt am Main, Berlin, Bern etc.

Rehm, G. (ed.) (2016b): *Language as a Data Type and Key Challenge for Big Data. Strategic Research and Innovation Agenda for the Multilingual Digital Single Market. Enabling the Multilingual Digital Single Market through technologies for translating, analysing, processing and curating natural language content*. July. Version 0.9. July 04, 2016. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded projects CRACKER and LT_Observatory.

Rehm, G. (ed.) (2017): *Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda*. December. Version 1.0. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded project CRACKER.

Rehm, G., S. Hegele (2018): "Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs". In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Schwenk, H. (2012) “Continuous Space Translation Models for Phrase-Based Statistical Machine Translation”. Proceedings of COLING 2012: Posters, pages 1071–1080, COLING 2012, Mumbai, December 2012.

Sennrich, R., B. Haddow, A. Birch (2016a): “Neural Machine Translation of Rare Words with Subword Units.” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1715–1725, Berlin, Germany, August 7-12, 2016

Sennrich, R., B. Haddow, A. Birch (2016b): “Improving Neural Machine Translation Models with Monolingual Data”. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 86-96).

Sennrich, R., O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli , A. Valerio Miceli Barone, J. Mokry, M. Nadejde: (2017): “Nematus: a Toolkit for Neural Machine Translation”. Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 65–68

STOA. (2017): *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March. http://www.europarl.europa.eu/stoa/.

Turchi, M., M. Negri, M.A. Farajian, M. Federico (2017): “Continuous learning from human post-edits for neural machine translation”. The Prague Bulletin of Mathematical Linguistics, 108(1), 233-244.

Williams, P., R. Sennrich, P. Koehn, M. Post (2016): “Syntax-based Statistical Machine Translation”. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers