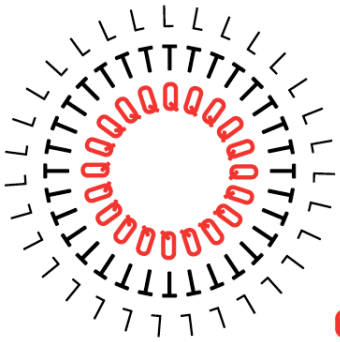
3.1 Publishable summary



qtleap

quality translation by deep language engineering approaches

Acronym and title
QTLeap- Quality Translation by Deep Language Engineering Approaches

Identification
QTLeap is a three year European **scientific and technology research project** on machine translation, planned for the period from November 1, 2013 to October 31, 2016.

Scope
Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language. Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses. There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

Executive summary
The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.
Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.
This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

QTLeap project explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

Motivation

In the last decade, mainstream research on Machine Translation (MT) has benefited mostly from the significant advances obtained with the exploitation of increasingly sophisticated statistical approaches. To a large extent, this incremental advancement has been obtained also by encompassing a host of subsidiary and increasingly more fine-grained linguistic distinctions that add to the surface level alignment on which these approaches are ultimately anchored.

It has been ventured in recent years, both in some leading academic and industry circles, that the incremental progress towards **quality MT** of this research path may be asymptotically reaching a ceiling. The claim relies on the observation that such curve tends to be more evident as more fine-grained distinctions are needed to aim at better translations with fewer gains in terms of quality increase.

Vision

MT systems based on **statistical shallow processing** have transformed language technology, taking advantage of the unprecedented amounts of data and processing power that has become available over the last twenty years to achieve goals that seemed unattainable before. To leap forward in terms of the quality of its output, machine translation technology will be moving beyond the current mainstream generation of statistical MT systems by taking advantage of enhanced capacities for deeper analysis of natural language and world knowledge that are now becoming available.

In the long run, high quality MT will rely on architectures based on rich knowledge approaches. Richer world knowledge will be available, even beyond the current Linked Open Data (LOD), with respect to larger datasets, richer semantics enhanced with world facts, and more dynamic conceptual knowledge representation.

Concomitantly, the evolutive trend in Natural Language Processing (NLP) shows a strong movement from knowledge-poor towards knowledge-rich language processing, supported by deep grammars and deep language resources.

These two streams of research — **rich world knowledge and deep linguistic knowledge** — will increasingly merge together, leveraging each other's potential, and the future in quality MT will belong to approaches increasingly based on deep language engineering approaches that rely on large-scale multilingual open data enriched with deep linguistic and semantic information.

Goal

The central goal of the present project is to obtain a **methodological advancement** in machine translation by pursuing a novel approach that breaks the way to higher quality MT and to a new cycle of technological advancement.

In this project, we build on the complementarities of the two pillars of NLP — symbolic and probabilistic — and seek a quantum leap in their hybridization. We explore combinations of them that amplify their strengths and mitigate their drawbacks by changing the angle from which this combination has been tackled in mainstream NLP in general, and in MT in particular, with a new design for the intertwining of statistical and rule-based approaches to MT.

Background

Deep language processing has its stepping stone in linguistically principled methods and, by embedding data based solutions, has been evolving towards supporting realistic natural language applications. In terms of efficiency, it has benefited from semantic underspecification techniques and advances in parsing technology that has fostered its speed. In terms of text coverage, it has

benefited from research on the automatic expansion of the lexicon. In terms of structural ambiguity resolution, word level processing have been outsourced to shallow pre-processing modules and language models have been coupled with grammars to rank their parses. And in terms of robustness, techniques have being developed to cope with multi-word expressions and to handle out of vocabulary words.
Importantly, as deep linguistic representation is in tight convergence with world knowledge representation, this progress can now be further supported also by building on the advances in terms of **lexical semantic processing** made possible by LOD and strengthened by enhanced referential and lexical ambiguity resolution techniques.
This line of steady progress has matured to a level that can now support a game-changing approach to MT. Concomitant key driving innovations — on both counts of algorithmics and language resources — have emerged recently that are further stepping stones for such a shift.

Approach
The construction of deep treebanks has progressed to be delivering now the first significant **multilingual Parallel DeepBanks**. In these datasets, pairs of synonymous sentences from different languages are annotated with their fully-fledged grammatical representations, up to the level of their semantic representation.
The construction of **LOD, ontologies and other semantic resources**, in turn, has also progressed now to be supporting impactful application of lexical semantic processing that handles and resolves referential and conceptual ambiguity.
These novelties are crucial in permitting for the cross-lingual alignment supporting translation to be established at the level of deeper linguistic and knowledge representation. The deeper the level of representation the less language-specific differences remain among source and target sentences and new chances of success become available for a statistically based transduction step.

Innovation
To exploit the potential of these new datasets and approach for a breakthrough in quality MT technology, new transduction algorithms have been emerging that seek to anchor their key **translation stage in deeper linguistic representations**. Initial experiments have delivered results that are highly competitive with regards to the top performing mainstream systems and that over perform the remaining competitors.
The full potentials of these initial results are waiting to be explored and unfolded by bringing into play further deep grammatical representations (e.g. abstract grammatical dependencies, tectogrammatical, minimal recursion semantics, etc.), further semantic collections (e.g. DBpedia, OpenCyc, etc.) and further advanced methods (e.g. Tree HMMs, Tree Kernels, etc.).

Aimed results
This project will deliver both an **articulated methodology for quality machine translation** that innovatively explores deep language engineering approaches to language technology, and an **empirically grounded validation** of its technological potential and impact.
In the pursuit of these results, this project will develop MT pilots delivering quality machine translation services whose exploitation will foster new solutions in the business sector, especially in all those vast areas for which machine translation of higher quality than current state of the art is still needed and in high demand.

Strategy
As the activities in the project will be deploying and the project will be moving forward along its timeline, we will be **moving towards quality MT based on progressively deeper language engineering approaches**. We will be moving to experimental settings where:

- the transduction steps are supported by more solutions for the lexical semantic processing and the resolution of ambiguity, starting with resolved names of entities and moving towards concepts isolated by means of word sense disambiguation transitively linked across different languages.
- the transduction steps are performed at deeper levels of semantic representation, starting from dependency representations at the syntax-semantic interface and moving towards resorting to increasingly more detailed and enriched semantic representations closer to logical forms.

Work plan

The project activities will be executed according to a work plan organized along two major and reciprocally supportive driving axes. As the activities in the project will be deploying and the project will be moving forward along its timeline, these activities will be progressing (i) towards quality MT based on increasingly deeper language engineering approaches, and (ii) towards validation and evaluation settings increasingly closer to a real usage scenario.

Thus the project work plan encompasses several milestones, which include the delivery of **four MT Pilot systems**, namely Pilot 0 (Milestone B, M6), Pilot 1 (Milestone D, M16), Pilot 2 (Milestone E, M24), Pilot 3 (Milestone F, M35)

The MT pilots to be constructed will be embedded in a multilingual call centre. This is a **real usage scenario** where high quality machine translation could not be called to play a more relevant and opportune role, to support efficiency and economy of scale, thus serving as a real life test bed for the extrinsic evaluation of the results to be achieved and for the validation of the project objectives.

Impacts

From a business and societal perspective, this project aims at producing a significant impact for **commercial quality machine translation**, the industry related to it, and for the European citizens, in general, as the ultimate users and beneficiaries of translation technology in their multilingual living and working environment.

From a R&D perspective, by means of its pilot advancement, this project aims at inducing a snowball effect that triggers and attracts the creation of more advanced algorithmics, deeper representations, larger volumes of data and a more multilingual and varied application environment: the ultimate goal is to give a decisive push towards opening a **new research cycle of sustainable progress in quality MT**, thus contributing for this project to "serve as a bridge to activities in Horizon 2020", the forthcoming EC's Framework Programme covering the period 2014-2020.

Consortium

The consortium undertaking this project is composed of partners that have a **longstanding research record** in MT and NLP, and are world leaders in pioneering the development of the algorithmics and the construction of the core systems and datasets underlying deep language engineering approaches to language technology.

They are bringing to the project datasets and systems that deliver grammatical representations of different levels of depth, from a range of different grammatical frameworks, and covering a range of languages from major European language families.

No.	Partner organisation name	Short name	Country
1	University of Lisbon, Faculty of Sciences	FCUL	Portugal
2	German Research Centre for Artificial Intelligence	DFKI	Germany
3	Charles University in Prague	CUNI	Czech Republic
4	Bulgarian Academy of Sciences	IICT-BAS	Bulgaria
5	Humboldt University of Berlin	UBER	Germany
6	University of Basque Country	UPV/EHU	Spain
7	University of Groningen	UG	The Netherlands
8	Higher Functions, Lda	HF	Portugal

Advisory Board

The direction of the project will be informed by the advice on strategic issues from the Advisory Board of Potential Users. This Advisory Board includes **industrial participants** that are ready to contribute with their advice on the strategic course of the project activities, and are interested in the innovation potential of the results targeted at by the project and will be in the first row of the potential users that will take the lead to exploit their business potential.

No.	Member organisation name	Short name	Country
1	CA Technologies Development Spain S.A.	CA	Spain
2	Eleka Ingeniaritza Linguistikoa SL	ELEKA	Spain
3	GridLine BV	GRIDLINE	The Netherlands
4	OMQ GmBH	OMQ	Germany
5	Ontotext AD	ONTOTEXT	Bulgaria
6	Lingea s.r.o.	LINGEA	Czech Republic
7	Seznam.cz, a.s.	SEZNAM	Czech Republic
8 (also partner)	Higher Functions, Lda	HF	Portugal

Approach to MT

QTLeap project counts on the contribution of partners that have, each of them, a wide range of strengths and backgrounds that, by bringing common and complementary technology, systems and resources, permit the project to embrace a common vision of undertaking research towards producing high-quality outbound MT by using more linguistic-intensive results.

This vision is being pursued under an approach shared among all partners of exploring deep language processing and of resorting to a common hybrid methodology that combines the best statistical and rule-based solutions.
To this end the partners have adopted a common architecture, based on transfer, and the same real-usage evaluation scenario. This scenario results from embedding machine translation into the workflow of online QA in ICT troubleshooting, from which a shared test dataset containing real users interactions was extracted and is being used by the partners.
The MT prototypes of the project are being developed along a progressive sequence of four pilots, where each one of these pilots covers every one of the seven languages in the project, in their translation pairs. The performance of these prototypes in this scenario is being assessed through a common set of evaluation metrics, thus ensuring full comparability of the results progressively obtained along the deployment of the project.
Summing up, all partners in the project share and are engaged with a common approach, namely:

- a common vision: to produce high-quality outbound MT using more linguistic-intensive results
- a common approach: deep processing;
- a common methodology: hybrid between rule-based and statistical;
- a common architecture: transfer-based;
- a common evaluation real-usage scenario: online QA in ICT trouble shooting;
- a common test dataset: interactions with users in the above real-usage scenario;
- a common set of evaluation metrics: automatic mainstream metrics suplemented with the multidimensional quality metrics;
- a common language (English) as target or source for each one of the seven languages in the project;
- a common path of progression for each language pair, ensuring comparability of the research exercise: every pair is developed along the four Pilots M0-M3

Results in reporting period
During the period reported in detail in the remainder of the present document, the project obtained the following results, which were planned in its work plan and are very briefly sketched as follows:
Documentation and dissemination (Milestone A; M2): The data sets and tools available to the consortium were fully documented and gathered and the forthcoming LRT curation work in the project was planned and put into action. The Strategic Advisory Board gathered in its first planned meeting to help fine tune the direction of the project. The website and all other dissemination instruments and plans were deployed.
Baseline MT Pilot 0 and pipelines (Milestone B, M6): The experimental workbench for developers to handle pipelines of tools and to develop MT tools based on state of the art SMT was set up. The state of the art MT Pilot 0 was developed and baselines for the rest of the project were obtained.
Application in real usage scenario (Milestone C, M12): The baseline MT Pilot 0 was integrated into the real usage scenario and a first extrinsic evaluation exercise was performed, with the resulting demonstrator permitting to use the industrial partner's QA system with the languages covered by the project.

Preparation of MT Pilots 1, 2 and 3 (forthcoming milestones, after reporting period): Preparatory work for the deployment of the MT Pilots 1, 2 and 3 was deployed according to plans.

Website
Updated information on the development of the project and its results can be checked at **http://qtleap.eu**