

Introducing the project “Centre for Next Generation Localisation”

Very often language and regional barriers restrain the free flow of information, products and services. The project “Centre for Next Generation Localisation” – an Academia-Industry partnership with over 100 researchers – helps to break down these barriers. It enables people to collaborate and disseminate their innovations and ideas without the barrier of language and cultural conventions.

About the project

The Centre for Next Generation Localisation (CNGL) project started in 2008 with a duration span of five years. It includes over 100 researchers developing novel technologies. Four Irish universities are involved in the CNGL: the Dublin City University, the Trinity College Dublin, the University College Dublin and the University of Limerick. The industry partners involved are Alchemy Software Development, DNP, IBM, Microsoft, SDL, SpeechStorm, Symantec, Translan and VistaTEC. There are also international collaborators from different universities throughout the world.

The director of the project, Professor Josef Van Genabith, describes the goal of CNGL: “Our work is guided by the vision of enabling people to interact with content, products and services in their own language, according to their own culture, and according to their own personal needs.”

Milestones

The project includes the following milestones:

- a) Integrated Language Technology
- b) Digital Content Management
- c) System Framework
- d) Localization

The research of each milestone is conducted at a different university. However, tutorials, seminars, demonstrations and conferences keep researchers informed about the latest news from the other universities. Let’s take a closer look at each milestone.

Integrated Language Technology

This research area includes improving the quality of Machine Translation (MT) and the deployment of MT into the localization workflow. Language service providers (LSPs) demand millions of automatic translations every day, making MT indispensable. MT started as being rule-based and moved to being more data-driven and machine-learning-based in the last years.

In general, researchers from the Dublin City University design, develop and evaluate MT engines for Next Generation Localisation. Furthermore, they include Speech Technologies (ST) in their research. Flexible and on-the-move voice access and response enhance the intelligent access of digital content.

Noteworthy is that various levels of speech are taken into account, such as voice characteristics, prosody and voice quality, emotional state or mood. The Integrated Language Technology track also includes the tasks of automatic annotation of localisation data with metadata and text classification in the field of Text Analytics (TA).

Digital Content Management

This area combines Adaptive Hypermedia and Information Retrieval (IR). The former – in combination with Adaptive Web systems – addresses the issue of personalization of hypermedia (WWW) documents in order to provide context-sensitive IR and reuse of digital content. The goal of Digital Content Management is to achieve advanced content retrieval and adaptive composition of multilingual digital content, so that the access to and navigation of multilingual corpora becomes easy. Multilingual Information Access (MLIA), cross-language information retrieval (CLIR), multimedia IR and web data sources are some of the areas researchers are working on.

System Framework

The System Framework research area looks upon improving the usability of Integrated Language Technology and Digital Content Management. Apart from the usability, it also takes into account human factors, work analysis, modality combination and cultural diversity. The evaluation of core technologies, the development of workflow management systems and information exchange mechanisms are some of the objectives of this area.

In short, System Framework assures that Integrated Language Technologies, Digital Content Management and Localization projects interoperate and are dynamically and flexibly integrated with user applications.

Localization

In the broadest sense, localization is the linguistic and cultural adaptation of digital content and services. It's main challenges are volume (the amount of content to be localized), access (on-the-move, instant access to digital content) and personalization (personalized information to individual user requirements).

This chapter is subcategorized into two sections: Localisation Research Centre (2.1) and CNGL and Localization" (2.2).

The Localisation Research Centre (LRC) is the information, research and educational centre for the localization industry. It is based at the University of Limerick (UL) since 1999 and its director is Reinhard Schäler. The LRC cooperates with worldwide digital publishers, the media, consultancy firms, government agencies, the European Commission and senior executives from the localization industry who are interested in future technologies and processes for globalization, internationalization, localization and translation (GILT).

It publishes an annual journal called "Localisation Focus - The International Journal of Localisation" which is the only peer reviewed and indexed academic journal focusing exclusively on localization and the localization industry. Furthermore, the LRC offers a Master of Global Computing and Localisation.

CNGL and Localization

There are eight Ph.D. students, three postdoctoral researchers as well as faculty members who work on localization in the terms of CNGL. Industrial collaborators include: DNP, Symantec, Alchemy Software Development, Translators Without Borders, TM-Europe, VistaTec, Microsoft, IBM. The international collaborators are from the University of Ottawa, the Universidade Federal de Santa Catarina, the Open University, U.K., the Queensland University of Technology and the University of South Africa.

The localization milestone is divided into three different work packages (LOC1, LOC2 and LOC3). Each postdoctoral fellow is responsible for one work package and supervises some Ph.D. students.

The LOC1 group works on the design and development of digital content guidelines and standards. One of the objectives is the development of the Localization Knowledge Repository (LKR), a web-service where digital content can be accessed, edited and evaluated. Digital Content is subcategorized into enterprise and personalized content – accordingly Enterprise Localization (EL) and Personalized Localization (PL).

The former is of high volume and for commercial reasons, as it caters for the needs of digital content publishers and service providers and the latter is of low volume, interactive and tied to Web 2.0 (wikis, blogs, social networking, etc.), addressing the often impromptu needs of the individual consumer of electronic content. Both kinds of EL and PL will be represented in the LKR.

LOC2 deals with the adaptation and evaluation of translation tools. It controls the automation of Translation Memory (TM) and Machine Translation (MT) systems as well as terminology databases (TDB). Its objective is to understand to which extent today's tools and technologies cover the requirements of the localizers, and to recommend better alternatives. The research group will try to develop a matrix for the measurement of the effectiveness, adaptability, usability and impact of digital content management technologies. Testing and Quality Assurance (QA) are also research areas of LOC2 group.

The LOC3 research group mainly works on the automation of workflow systems. It specifies localization scenarios with their associated workflow ranging from bulk localization to the personalized production of digital information and ad hoc social networking. Quality Management is also researched by the LOC3 group.

In summary, the localization research groups design and develop the digital content

(LOC1), translate and adapt it (LOC2) and automate and standardize the workflow of the localization factory (LOC3).

The main goal of the localization milestone with regard to CNGL is to lead the way toward Next Generation Localization, after combining the results of Integrated Language Technology and Digital Content Management. The instant, on-the-fly service should have quality, which is based on linguistic, cultural and targeted personalization of digital content. Users' requirements and preferences are taken into consideration and cultural, linguistic and socio-economic barriers are removed.

Conclusion

The CNGL started in 2008 with the aim to achieve the "Next Generation Localization". Integrated Language Technology, Digital Content Management, Localization and System Framework are the interrelated research tracks of the CNGL project. The first two are basic research tracks only, whereas the remaining are applied and integrating research tracks. The main goal, which is the Next Generation Localization, covers both Enterprise and Personalized Localization.