http://www.cosyne.eu

CoSyne: Annual public report 2012

Distribution: Public

CoSyne

Multilingual Content Synchronization with Wikis

FP7-ICT-4-248531 Deliverable

Version 1.0, November 22, 2012

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development.

This report summarizes the work performed within the context of the CoSyne project through its second and first part of the third year.

1 Project description

Wikis have gained increasing popularity over the last few years as a means of collaborative content creation, as they allow users to set up and edit web pages directly. A growing number of organizations use Wikis as an efficient means of (collaboratively) providing and maintaining information across several sites.

Currently, multilingual Wikis rely on users to manually translate different Wiki pages on the same subject. This is not only a time-consuming process, but also the source of many inconsistencies. This is because users update the different language versions separately, and every update would require translators to compare the different language versions and synchronize the updates. The overall aim of the CoSyne project is to use statistical machine translation techniques, structural analysis techniques and entailment techniques to automate the dynamic multilingual synchronization process of Wikis.

2 Summary of activities

The period of March 1, 2012 to September 1, 2012, focused on evaluating the second prototype delivered at the end of year 2, and further developing the system in preparation for the next delivery of the prototype at the end of year 3.

In the first two years, the focus was on English, German, Italian and Dutch. This year (year three), we address the adaptability of the CoSyne architecture to languages with limited resources. To this end, two additional languages from different language families are added (Bulgarian and Turkish). One of the purposes of adding two languages (Turkish and Bulgarian) in year three is to measure to what extent the CoSyne system can be adapted to novel languages with more limited resources and how much effort is required for this.

In the previous year, components of CoSyne were integrated through web services with the Media Wiki platform. Subsequent testing by end-users revealed issues with stability and resource use of the system, these have now been resolved by optimizing various components and distributing the computation load over several machines. Furthermore, a component has been developed that learns from users' corrections to improve translation quality.

The consortium has put the CoSyne system online. It is available at: `http://prototype.cosyne.eu/demo/en/index.php/Main_Page`.

The following sections provide a more detailed overview of the activities, grouped by work package.

3 Robust dynamic Machine Translation (MT)

The objective is to develop a robust machine translation component and integrate this into a multilingual Wiki content management system.

In the previous year, the lack of resources for certain CoSyne language pairs was compensated for by combining translation models for other pairs. Over the last year, additional resources have been harvested and combined with the previous models for the year 1 languages English, German, Italian and Dutch. The resulting systems have been integrated with the online version of the CoSyne system.

In this third year we are in the process of adapting the system to Turkish and Bulgarian.

4 Cross-lingual content entailment

The purpose of this module is to identify textual content overlap between segments of Wiki pages across languages, in order to avoid redundant machine translation. For this purpose the semantic analysis techniques have been implemented that are necessary to synchronise the content of Wiki pages about the same topic but written in different languages.

Textual Entailment (TE) recognition is the core technology for performing such annotations that determine which portions of the input pages have to be translated for their mutual update in the subsequent steps of the synchronisation process. We identify the optimal insertion points for translated content in order to preserve coherence.

In the past year, activity on Cross Lingual Textual Entailment focused on using word alignment information as a mechanism to detect multidirectional entailment relations. In order to address language portability issues, the approach has been tested in two different conditions: i) when both parallel and annotated CLTE data are available, and ii) when only parallel data are available. In the latter (harder) case, CLTE models are re-used for language pairs for which CLTE annotated pairs are missing. Additional effort has also gone into finalizing the first round of a CLTE task within the SemEval 2012 evaluation campaign, and on organizing the second round which will take place next year.

5 Adaptive and self-learning MT

The objective is to develop a robust machine translation component for six languages that is fully integrated into a multilingual Wiki content-management system. The MT component will be able to translate user-generated content with sufficient quality by handling noisy input that can contain typos and ungrammatical sentences, and adapt to the category of the source document. In addition, the MT component will interact with the user edits dynamically by translating edited parts and inserting them in the right places, leaving as much of the original target side intact as possible.

In the reporting period, test sessions with professional editors were organized. These have yielded user corrections of machine translation output for the language pair German-English. Due to practical limitations, only a limited number of corrections was obtained. In practice, this data was not enough to train a useful self-learning correction system on. Therefore, an additional data set of simulated user edits was created which was based on corrections extracted from the edit history of Wikipedia pages. The resulting correction system seems to perform better, and user evaluation is currently underway.

6 Language-independent induction of structure for Wikis

The objective is to introduce, transfer and adapt the structure of Wiki pages. In this context, structure means text segmentation into sections, detection of insertion points for newly translated content, cross-lingual alignment of segments, hyperlinks to other Wikis and info boxes - short lists in attribute-value format that capture particularly salient information about the topic of the Wiki to which they belong.

In the reporting period, further work was done on a cross-lingual segmentation and alignment module. A probabilistic method for aligning text segments across documents and across languages was used which is based on induced topics. To bridge between languages and documents and to build multilingual topic models, terms in a document are linked to an inventory of concepts derived from Wikipedia.

A comparison to other state-of-the-art approaches has shown that this new segmentation and alignment module is competitive.

7 MT usability evaluation

The objective is to evaluate the quality of the translations produced by the system. This is an ongoing task inasmuch as it permits feedback to the MT development work packages. Likewise, the end-user evaluation has two phases, the first of which can be seen as a pilot.

The evaluation activities all belong to one of three categories:

1. general translation quality,
2. fine-grained diagnostic evaluation to help identify areas where MT systems can be improved,
3. adequacy of translation for end-user.

In the reporting period, several test sessions were organized. A toolkit for the diagnostic evaluations called DELiC4MT was developed in an earlier phase of the project, and its functionality has been extended over the course of the last year.

8 Demonstrator

At the end of the second year, the first prototype was delivered and shown to the EC. As a result of last year's activities, the second version of the protoype was delivered. Improvements include a friendlier user interface and visualization of its internal workings for debugging purposes. In total, three versions will be built, one for each year of the project, to serve as demonstration integrated prototypes. The third integrated prototype will also be used as the project's Showcase/Demonstrator.

The current prototype (see Figure 1 for a screenshot) covers German, English, Dutch and Italian. The third prototype will also cover Bulgarian and Turkish. The code of all prototypes will be made available on SourceForge, under a public license. Integration of the technology in these demonstration environments will be done in collaboration with experts from the Wikimedia Foundation (the Dutch chapter, and possibly others as well).

The users have specified the requirements of the system, for the professional user group: journalists, editors, wikipedia contributors, etc. Specific user requirements discussed include user-friendliness; focus on language, text and content; interfacing requiring no programming skills; search capability, spell checker, etc. A use case is specified for the use of the system in Kalenderblatt/Today in History (both Deutsche Welle sites).

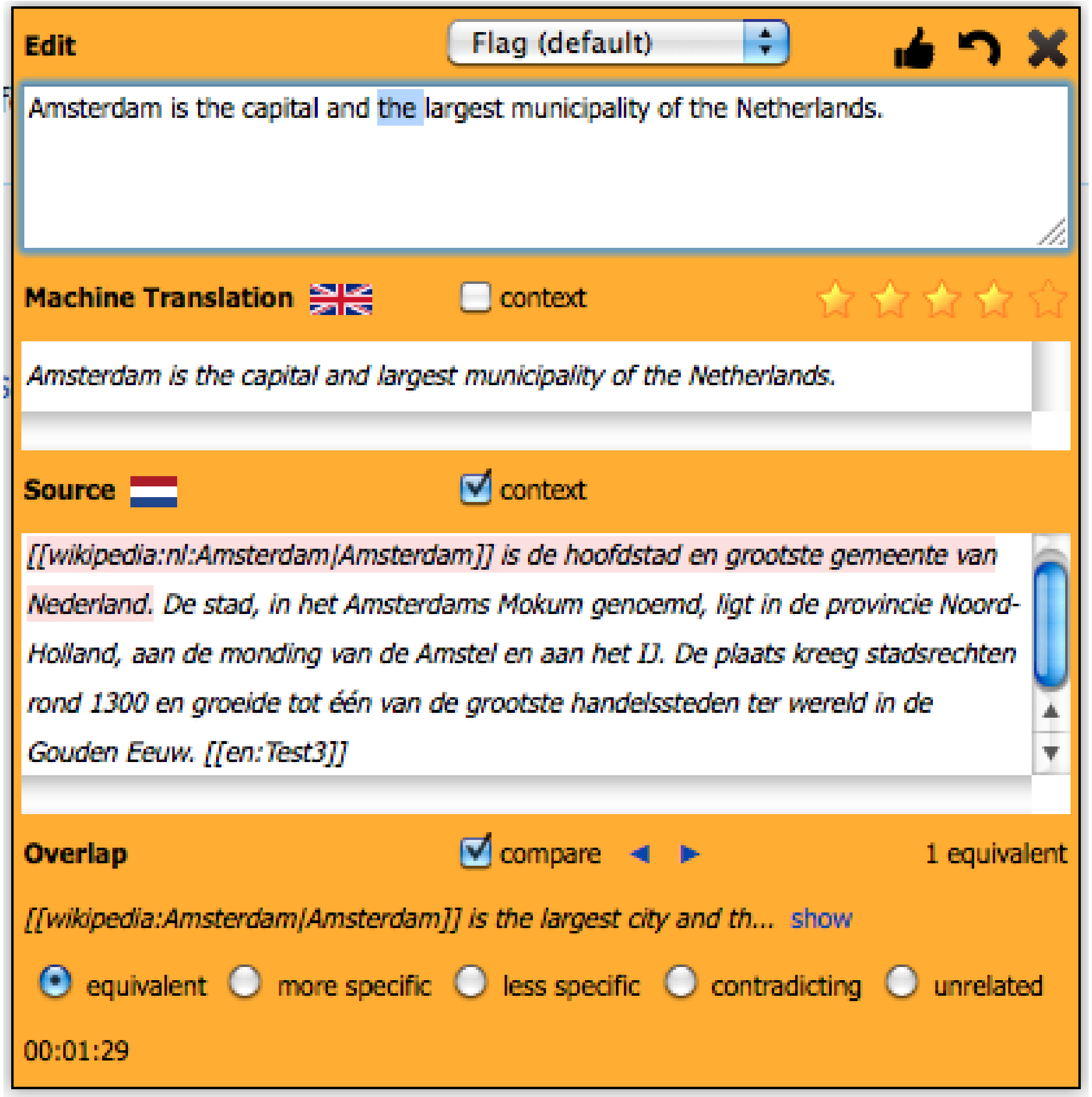
9 User involvement, dissemination and exploitation

During this reporting period exploitation was a key issue of discussion and planning, especially as results have become visible and the project is nearing its end. The deployment and evaluation of the different use cases for the three user partners, Deutsche Welle, Netherlands Institute for Sound and Vision (NISV) and Wikimedia Foundation Netherlands, give a clear direction towards the usability and exploitability of the system.

The entire consortium, i.e. technical partners (component developers), integrator, and user partners, have discussed and outlined the various possibilities to exploit the system after the conclusion of the project. This relates to the integrated system as well as to the separate components.The different options and their business potential are considered.

Exploitation is also closely related to the increased dissemination in this last period, to show potential customers and/or users what the system is capable of. Over the last 12 months, scientific papers were presented at venues such as MT Marathon2012, Edinburgh and WikiSym 2012. A comprehensive list can be found at `http://cosyne.eu/index.php/Publications`.

Among the presentations and demos were:



Edit Flag (default)

Amsterdam is the capital and the largest municipality of the Netherlands.

Machine Translation ☐ context ★★★★☆

Amsterdam is the capital and largest municipality of the Netherlands.

Source ☑ context

[[wikipedia:nl:Amsterdam|Amsterdam]] is de hoofdstad en grootste gemeente van Nederland. De stad, in het Amsterdams Mokum genoemd, ligt in de provincie Noord-Holland, aan de monding van de Amstel en aan het IJ. De plaats kreeg stadsrechten rond 1300 en groeide tot één van de grootste handelssteden ter wereld in de Gouden Eeuw. [[en:Test3]]

Overlap ☑ compare ◀ ▶ 1 equivalent

[[wikipedia:Amsterdam|Amsterdam]] is the largest city and th... show

● equivalent ○ more specific ○ less specific ○ contradicting ○ unrelated

00:01:29

Figure 1: Screenshot of the CoSyne system in action.

- setting up an online demo of CoSyne (up & running since August 2012, publicly open)
- WikiSym2012, August 2012, Austria, Linz (demo)
- Presentation and demo at Wikizaterdag, a monthly gathering of Dutch Wikimedia members that was dedicated to CoSyne (September 2012, Utrecht Netherlands)
- Presentation and demo at the Software Freedom Day 2012 (September 2012, Amsterdam, Netherlands)
- Presentation at WCN, the Dutch Wikimedia conference (November 2012, Utrecht Netherlands)

In terms of collaboration, in October we had a meeting with the Wikimedia Foundation - German Chapter to explore the possibility of working together and to see if they can participate in the evaluation of the system. They are interested in evaluating the system, and they will be participating in the Final Showcase meeting of CoSyne.

Project material was further developed and actively disseminated during conferences and fairs attended by the project partners. The CoSyne project wiki was continuously updated, a new CoSyne poster was produced and a new CoSyne video was released by DW (see Figure 2), which explains the purpose and set-up of the CoSyne project. It can be found on YouTube and on `http://cosyne.eu/`.

Figure 2: Screenshot from demo movie.

10 Project Status

CoSyne is meeting its planned goals and objectives.

11 Further Information

For further information see: `http://www.cosyne.eu`.
For further information please contact Christof Monz, University of Amsterdam, `c.monz@uva.nl`.