http://www.cosyne.eu

CoSyne: Annual public report 2011

Distribution: Public

CoSyne

Multilingual Content Synchronization with Wikis

FP7-ICT-4-248531 Deliverable

Version 1.0, January 9, 2012

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

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 automate the dynamic multilingual synchronization process of Wikis.

2 Summary of activities

The focus of the CoSyne project in 2011 was on setting up the infrastructure, clarifying the interfaces, and building the initial set of models within the various work packages. The components of CoSyne have been integrated through web services with the popular, open-source Media Wiki platform.

Guided by the end-users, a number of datasets have been identified which will be used throughout the project to evaluate and measure progress.

By producing the architectural specifications and end-user requirements it became much clearer to all partners involved in the construction of the prototype exactly how the components will interact. The end-user feedback also further clarified how the system is to be exploited in a real-world setting. This helped the partners to calibrate the development of the anticipated functionalities of the prototype.

In addition to the test sets used for the overall evaluation of the system, a collection of test sets for assessing the quality of the machine translation component has been produced, and commercial translation systems have been evaluated against it by means of a spectrum of evaluation metrics. We started with creating the benchmark sets for evaluating the textual entailment component.

What follows is 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. The MT component will be able to handle typos and ungrammatical sentences. In addition, the MT component will dynamically interact with the user edits by translating edited parts and inserting them at the appropriate places in the target document, leaving as much of the original target side intact as possible.

All required resources have been harvested by crawling multilingual websites, and lack of resources for some language pairs has been compensated for by combining translation models. This has made it possible to build initial translation systems for the year 1 languages English, german, Italian and Dutch. An online version of our statistical machine translation system has been implemented which achieves robust translation performance.

4 Cross-lingual content entailment

The objective 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. The task will annotate the pages in terms of fully overlapping, partially overlapping, and non-overlapping segments. 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.

A benchmark (dataset) for cross-lingual textual entailment has been developed and released to the public. It consists of a large number of (XML) annotated Text/Hypothesis pairs for the language combinations English/Italian, English/German, and German/Italian, as well as monolingual sets for English/English and Italian/Italian.

A new version (2.1) of the EDITS package for monolingual TE has been released this year. A baseline system for cross-lingual TE has been developed by combining MT with monolingual TE. This approach is modular and can thus easily benefit from improvements in both TE and MT. A different approach is to more tightly integrate the MT and TE techniques, which allows for more control over the system's behaviour. Experiments in this direction have been conducted utilizing distinct sources of multilingual lexical knowledge. The (published) results indicate the potential of this approach.

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 in a dynamic way by translating edited parts and inserting them in the right places, leaving as much of the original target side intact as possible. The adaptive aspect will improve translation quality by incorporating information about the domain of a document.

Recently, a system has been developed which can automatically distinguish between factual changes and translation corrections. This classifier is used to determine which parts of an updated Wiki page to translate. Also, the changes it labels as translation corrections will be used as input for a statistical correction model. The user corrections are exploited to learn from the translation mistakes and improve the translation model for future translations.

6 Language-independent induction of structure for Wikis

The objective is to introduce, transfer and adapt the structure of Wiki pages. By structure we mean text segmentation into sections, hyperlinks to other Wikis from the users' sites, and info boxes - short lists in attribute-value format that capture particularly salient information about the topic of the Wiki to which they belong. This involves, among other things, disambiguation of polysemous words and detection of topic changes, which indicate section boundaries.

A hyperlink identification module, based on a concept inventory extracted from Wikipedia, has been integrated into the prototype.

A cross-lingual coarse alignment has been integrated as well. It uses output from the concept identification module to find overlapping text fragments in pages for different languages, and is thus language independent. A non-parametric probabilistic model is currently developed for English, and we are investigating methods for transforming this into a cross-language model.

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.

8 Demonstrator

The first prototype has been delivered and shown to the EC. We are currently in the process of building the second version of the protoype. 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. Prior to the software release, the architecture specification and user requirements were established. MediaWiki, as one of the most dominant platforms in use, is the basis of the CoSyne system. CoSyne is based on a Web Services architecture, with four computing components and the web interface accessible over a network. The project follows the Boehm spiral life-cycle model, allowing for iterative evolution of the components. The first two prototypes cover 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 the GNU General Public License. Integration of the technology in these demonstration environments will be done in collaboration with experts from the Wikimedia Foundation.

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, promotion, and awareness

The three end-user partners of the consortium, Deutsche Welle, Wikimedia Foundation Netherlands, Nederlands Instituut voor Beeld en Geluid, will deploy, integrate into their daily workflow, and evaluate the CoSyne system, which will give a clear direction towards the exploitability of the project's outcomes.

We have made good progress with respect to the dissemination. The project website, in Mediawiki format, was developed, set up and launched. This wikisite, providing content synchronisation with wikis, contains a public part (including publications, software and datasets) for dissemination purposes and a restricted (private) section which is accessible only to the project partners and used as a joint collaboration tool. Flyers and poster, all in English, were created. This material has been actively disseminated during conferences and fairs attended by the project partners.

In the period of the last 18 months (July 2010 up to January 2011), scientific papers were published in:

- Proceedings of ACL-2010, 48th Annual Meeting of the Association for Computational Linguistics,
- Proceedings of the Third Joint EM+/CNGL Workshop ”Bringing MT to the User: Research Meets Translators”,
- Proceedings of Machine Translation Summit XII,
- Proceedings of the TextInfer 2011 Workshop on Textual Entailment,

- Proceedings of EMNLP-2011, Conference on Empirical Methods in Natural Language Processing,
- Proceedings of ACL-2011, 49th Annual Meeting of the Association for Computational Linguistics,
- Proceedings of RTE-6 2010, Recognizing Textual Entailment Challenge (RTE-6) at TAC 2010,
- Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011).

There were presentations at:

- 10th DIVERSE Conference (NISV),
- AAAI 2010: 24th AAAI Conference on Artificial Intelligence (HITS),
- ACL 2010: 48th Annual Meeting of the Association for Computational Linguistics (UvA and FBK),
- Joint EuroMatrix-CNGL, JEC Workshop 2011 (DCU),
- MT Summit XIII (DCU),
- Wikimania 2011 (UvA, DW),
- EMNLP-2011 (FBK),
- ACL-HLT 2011 (FBK),
- CoSyne Promo Video (DW),
- EAMT 2011 (DCU, DW),
- META-NET Media and Information Services Vision Group Meeting 2011 (DW).

10 Future work

We are currently in the process Adding translation capabilities for Bulgarian and Turkish is among the activities planned for 2012. The quality of existing translation capabilities will be boosted by acquiring and training additional linguistic resources. Apart from this, enhanced user interface and support of additional scenarios are part of the development plan for the next prototype versions.

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`.