CoSyne: Synchronizing Multilingual Wiki Content

Amit Bronner
Netherlands Institute for Sound and Vision
Media Park, Hilversum,
The Netherlands
abronner@beeldengeluid.nl

Matteo Negri
FBK
Trento, Italy
negri@fbk.eu

Angela Fahrni
HITS gGmbH
Heidelberg, Germany
Angela.Fahrni@h-its.org

Yashar Mehdad
FBK
Trento, Italy
mehdad@fbk.eu

Christof Monz
University of Amsterdam
Amsterdam, The Netherlands
c.monz@uva.nl

Abstract
CoSyne is a content synchronization system for assisting users and organizations involved in the maintenance of multilingual wikis. The system allows users to explore the diversity of multilingual content using a monolingual view. It provides suggestions for content modification based on additional or more specific information found in other language versions, and enables seamless integration of automatically translated sentences while giving users the flexibility to edit, correct and control eventual changes to the wiki page. To support these tasks, CoSyne employs state-of-the-art machine translation and natural language processing techniques.

Author Keywords
Multilingual Content Synchronization, Cross-lingual Topical Alignment, User Edits Classification, Cross-lingual Textual Entailment, Context-sensitive Machine Translation, Wiki, MediaWiki, User generated content, RESTful web services

ACM Classification Keywords
H.3.5 [Online Information Services]: Data sharing; I.2.7 [Natural Language Processing]: Machine translation, Text analysis; I.2.1 [Application and Expert Systems]: Office automation

Introduction

CoSyne[7] is a European research project that develops a system for synchronizing the content of wiki pages across multiple languages. The system allows users who read a wiki page in one language to view overlapping and non-overlapping information automatically translated from linked pages in other languages. Additional information is displayed at the place that best maintains the coherence of the page. More specific information is displayed as alternative to existing content. The system highlights the sentences that it suggests to add or replace, and the user may accept, ignore or correct these suggestions.

The CoSyne system is a distributed web application built using REST-style architecture. A central server handles user requests, provides the business logic, and exposes RESTful interfaces for the transfer of resources between the following components:

1. Structural analysis of wiki pages including concept identification, cross-lingual topical alignment and the identification of insertion points for new information [2, 8];
2. Classification of user edits from page revision histories as factual edits or fluency edits [1];
3. Cross-lingual textual entailment in order to determine whether content is equivalent, more specific, less specific or unrelated [3, 4, 5];
4. Context-sensitive machine translation for translating text in a way that takes existing context into account [6, 9].

The system is deployed at different research centers as illustrated by Figure 1. The project languages include Bulgarian, Dutch, English, German, Italian and Turkish.

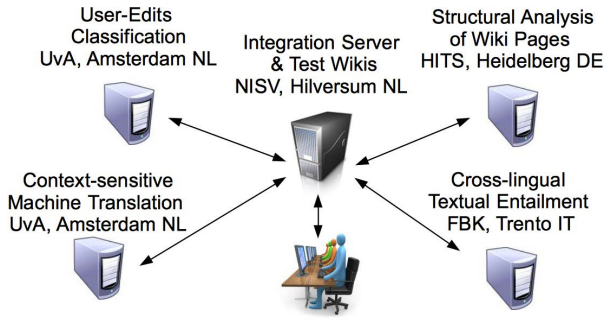

Figure 1: CoSyne distributed web application

For wiki users, CoSyne is a gadget[1] that could be activated via the user's preferences page. The gadget communicates with the web application and provides the following features:

- *Explore*: per-sentence display of overlapping and non-overlapping information from other language versions;
- *Suggest*: highlighting of insertion points for additional information and sentences that may be replaced by more specific information from other language versions;
- *Edit*: seamless integration of automatically translated sentences from other language versions, flexible editing and correction by users, standard preview and saving of changes to the wiki page.

[1] http://www.mediawiki.org/wiki/Extension:Gadgets

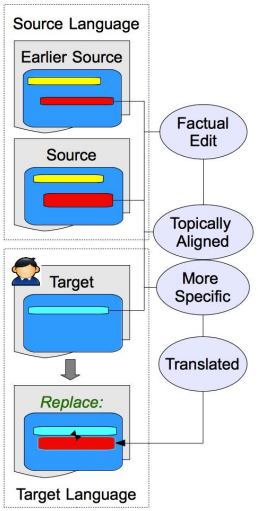

Figure 2: Synchronization example: a factual edit on the source page (red color) is topically aligned with text on the target page (cyan color). The source text is more specific, therefore the system suggests to replace the target text with a context-sensitive translation of the source text. No suggestion is made for the non-factual edit (yellow color).

The Synchronization Process

The system considers three page revisions as input:

- *Target*: the page that the user reads;
- *Source*: a linked page written in another language;
- *Earlier source*: optionally, an earlier revision of the *source* page, relevant when users synchronize pages on a regular basis.

The structural analysis component identifies concepts and topics and uses them to align sets of source sentences with sets of target sentences. It also provides an ordering of the sentences which is used for selecting insertion points in the target page.

If an earlier source revision is present, the user edits classification component compares the two source revisions, identifies sentence-level user edits and classifies them as factual edits or fluency edits.

Source sentences that contain factual changes with respect to the earlier source revision are paired with topically-aligned target sentences. The cross-lingual textual entailment component identifies multi-directional entailment relations between these sentences. This allows the system to determine whether a source sentence is equivalent, more specific, less specific or unrelated to a target sentence.

The context-sensitive machine translation component translates source sentences into the target language. It takes into account pre-edited sentences in the earlier source revision and existing target sentences in order to provide the most suitable translation in the given context.

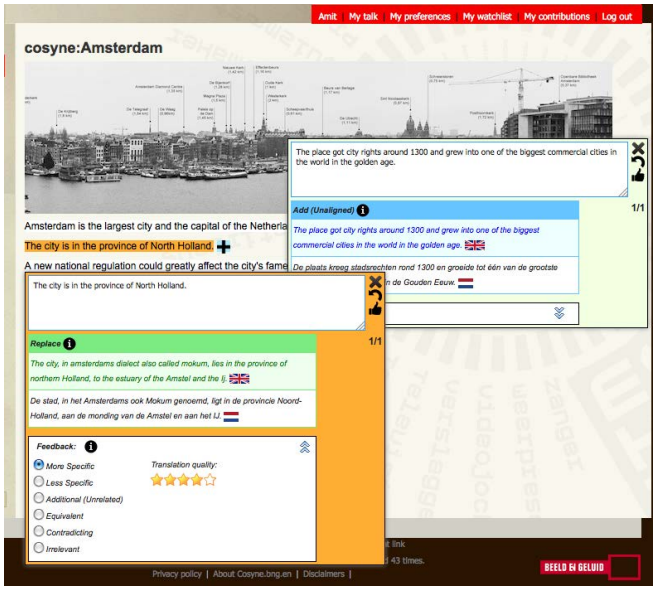

Figure 3: Demo screenshot: an English wiki page with a linked Dutch version. On the left (green color), CoSyne suggests to replace an English sentence with a translated Dutch sentence that contains more specific information. On the right (blue color), CoSyne suggests to add a translated Dutch sentence that contains additional information.

Finally, CoSyne provides a set of suggestions which is based on the output of the different components for all source and target sentences. The system may suggest to add a translated source sentence at a particular place on the page, to replace a target sentence with a translated source sentence, or to keep a target sentence as is. Figure 2 illustrates a simple synchronization example.

Demo Outline

The demo consists of two parts. The first part introduces the CoSyne system, its capabilities and features. A simple

and short example is used for initial demonstration. In this example, a user reads a wiki page in English and asks CoSyne for additional information from a linked wiki page in Dutch. The text contains a Dutch sentence which is equivalent to an English sentence, a Dutch sentence which is more specific than an English sentence, a Dutch sentence which is unrelated to any English sentence and an English sentence which is unrelated to any Dutch sentence. Figure 3 is a screenshot from the initial demonstration.

The second part of the demo presents more complex synchronization examples taken from wiki sites of the project's end-user partners:

- Deutsche Welle's "Kalenderblatt" ("Today in History"): a German-English website providing historical information.[2]
- Sound and Vision Collection Wiki: a Dutch website providing background documentation on television productions, actors, directors and news topics.[3]

Users can interact with the system during the demo: invoke the synchronization process and explore its results; integrate translated sentences, edit and correct them; preview the modified version and save changes to the wiki; and provide online feedback on system suggestions and on translation quality.

Conclusion

CoSyne addresses the challenging task of multilingual synchronization of user-generated content. The demo showcases the state of the art as well as the complexity of the task. The user interface allows for interactive exploration of the system during the demo.

[2]Deutsche Welle is Germany's international broadcaster providing news content in 30 languages.

[3]Netherlands Institute for Sound and Vision is one of Europe's largest audiovisual archives.

Acknowledgements

The research is co-funded by the European Commission through the CoSyne project FP7-ICT-4-248531.

References

[1] Bronner, A., and Monz, C. User Edits Classification Using Document Revision Histories. In *Proc. EACL* (2012).

[2] Fahrni, A., Nastase, V., and Strube, M. HITS' Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task. In *Proc. NTCIR-9* (2011).

[3] Mehdad, Y., Negri, M., and Federico, M. Towards Cross-lingual Textual Entailment. In *Proc. NAACL HLT* (2010), 321–324.

[4] Mehdad, Y., Negri, M., and Federico, M. Using Blingual Parallel Corpora for Cross-lingual Textual Entailment. In *Proc. ACL HLT* (2011).

[5] Mehdad, Y., Negri, M., and Federico, M. Detecting Semantic Equivalence and Information Disparity in Cross-lingual Documents. In *Proc. ACL* (2012).

[6] Monz, C. Statistical Machine Translation with Local Language Models. In *Proc. EMNLP* (2011), 869–879.

[7] Monz, C., Nastase, V., Negri, M., Fahrni, A., Mehdad, Y., and Strube, M. CoSyne: a Framework for Multilingual Content Synchronization of Wikis. In *Proc. WikiSym* (2011), 217–218.

[8] Nastase, V., Strube, M., Brschinger, B., Zirn, C., and Elghafari, A. WikiNet: a Very Large Scale Multl-ingual Concept Network. In *Proc. LREC* (2010).

[9] Yahyaei, S., and Monz, C. Decoding by Dynamic Chunking for Statistical Machine Translation. In *Proc. MT Summit XII* (2009), 160–167.