CoSyne: A Framework for Multilingual Content Synchronization of Wikis

[Extended Abstract]

Christof Monz[1], Vivi Nastase[2], Matteo Negri[3], Angela Fahrni[2], Yashar Mehdad[3], Michael Strube[2]
[1]University of Amsterdam, Amsterdam, The Netherlands
[2]HITS gGmbH, Heidelberg, Germany
[3]FBK-irst, Trento, Italy

ABSTRACT

Wikis allow a large base of contributors easy access to shared content, and freedom in editing it. One of the side-effects of this freedom was the emergence of parallel and independently evolving versions in a variety of languages, reflecting the multilingual background of the pool of contributors. For the Wiki to properly represent the user-added content, this should be fully available in all its languages. Working on parallel Wikis in several European languages, we investigate the possibility to "synchronize" different language versions of the same document, by: *i)* pinpointing topically related pieces of information in the different languages, *ii)* identifying information that is missing or less detailed in one of the two versions, *iii)* translating this in the appropriate language, *iv)* inserting it in the appropriate place. Progress along such directions will allow users to share more easily content across language boundaries.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Machine translation; I.2.7 [**Natural Language Processing**]: Text analysis; I.2.11 [**Distributed Artificial Intelligence**]: Languages and structures

General Terms

Wiki, multilinguality, translation, recognizing textual entailment

1. INTRODUCTION

At the core of Web 2.0 is the user as a content co-creator, as a result of which many of the Web 2.0 sites are written in languages other than English. For example, Wikipedia now features multiple language versions, 36 of which have more than 100,000 pages each[1]. While large amounts of information are already present in one language, users often start from scratch setting up a new language version for the same subject. This does not only introduce redundant efforts but also introduces sources of inconsistency between the different language versions of a Wiki page. While it is valuable to have diversity of opinion and content across the different languages, it prohibits a common information space across the languages.

Previous efforts in supporting multilingual Wikis focused on providing users with tools that facilitate collaborative content translation [1], or with a machine translation system embedded in the Wiki environment that provides full page translations, to be added as proper pages and edited by users [2]. In contrast to these approaches, CoSyne – a three year EU project (http://www.cosyne.eu/) – gives precedence to the existing manually produced content while synchronizing Wiki pages in multiple languages and maintaining consistency while the pages are edited by users in their preferred language. The processing consists in automatically merging pages in different languages by identifying the text fragments that are not shared by the different language versions, translating and embedding these fragments where appropriate. In this paper we present the challenges we have identified in this task, the approach we have adopted, the steps required to achieve our goals and the progress made so far.

2. THE COSYNE SYSTEM

The CoSyne system identifies content discrepancies across different language versions of Wiki pages, merging them to produce synchronized versions. Figure 1 shows a schematic representation of the system. Processing a Wiki page proceeds through the system components, each of which adds one or more layers of annotation, building upon the previously gathered information. The end result of the annotation process is used to produce the latest version of the Wiki, whose content reflects now the merged pool of information from all language versions. The system's architecture consists of the following components:

Structural Analysis analyzes the structure of the input Wiki pages, automatically identifying and categorizing segments that represent semantically coherent chunks (*e.g.* the A and B text fragments in the input Wiki documents in Figure 1). This processing relies on recognizing concepts mentioned in the text relative to a large-scale multilingual concept network [7]. Structural analysis allows to: *i)* preserve link structures across Wiki pages in different languages, *ii)* support the Content Entailment component by reducing the search space of overlapping/non-overlapping information, and *iii)* support the Adaptive and Self-Learning MT Correction component by providing segment categories that are used for domain adaptation.

Entailment-based Content Merging is in charge of annotating the input pages in terms of: *i)* overlapping information that does not need to be translated for synchronization, and *ii)* information that has to be translated

[1]`http://meta.wikimedia.org/wiki/List_of_Wikipedias`; April 4^{th}, 2011.

WikiSym'11, Oct 03–05 2011, Mountain View, CA, USA
ACM ACM 978-1-4503-0909-7/11/10.

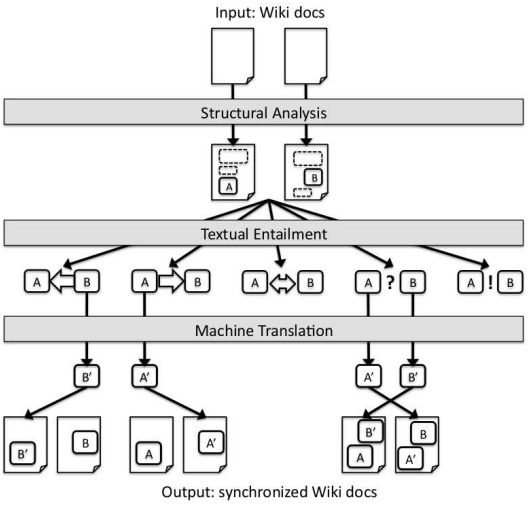

Figure 1: The automatic content synchronization process.

and has to migrate from one page to the other (*i.e.* more specific information, or factual information that is present only in one page). Such process is based on determining (uni/bi-directional) entailment relations between topically-related text fragments, like A and B in Figure 1). The output of this component allows the MT component to focus on translating content that is novel with respect to the Wiki page into which translated content is to be inserted. Recent work along this direction addressed the possible approaches to cross-lingual textual entailment [3, 4], and the collection of cross-lingual textual entailment data [5, 6].

Machine Translation translates the novel content following a statistical machine translation approach. The machine translation component pays particular attention to the issue of robustness and the translation of dynamic and user-generated content. It's *self-learning* subcomponent analyzes user edits and classifies them as either factual changes or translation corrections. Factual changes are dealt with by the machine translation component, while translation corrections are used to adapt the translation models and thereby improve translation quality. The performance of the MT component is monitored by an evaluation module, whose findings are used to improve the MT proper and the self-learning component.

Web Service The interaction between the users and the components are realized through web services that are integrated within the open source MediaWiki software. The Web service mediates the exchange and aggregation of the annotation layers produced by each component.

The final version of the system will be deployed and evaluated by the end user partners of the project:

- Deutsche Welle's Kalenderblatt/Today in History – Deutsche Welle (DW) is Germany's international broadcaster providing news content in 30 languages. Kalenderblatt/Today in History are web sites in German and English providing historically relevant information. They are independently generated, but cover the same events/themes. Automatic content syncronization would increase the efficiency of news editing.
- Sound and Vision Collection Wiki – Sound and Vision (NISV) is one of Europe's largest audiovisual archives. It maintains a Wiki that provides background text documents on television productions, actors, directors and news topics. Expanding access to this Wiki to a varied contributors/reader base will allow content from other sources to be integrated into the Dutch pages.

At the end of the project, the CoSyne system will cover the following languages: Bulgarian, Dutch, English, German, Italian and Turkish.

3. CONCLUSIONS AND FUTURE WORK

In this paper we described the main challenges addressed by CoSyne – a EU funded project (FP7-ICT-4-24853) for synchronizing the content of parallel multilingual Wikis. The project's aim is to provide an automatic way of bridging content across language boundaries, not by simply translating pages from one language to another, but by automatically analysing and mutually updating existing pages in different languages. So far, the overall approach has been defined, and several core components have been developed (*i.e.* structural analysis, entailment-based content merging, and MT components) and integrated. Several interesting issues, however, still have to be tackled. Apart from the extension to new language pairs, these include *i)* the identification and proper management of contradictions found in the input pages, and *ii)* the development of improved techniques for identifying the most appropriate entry points for translated information that migrates across pages.

Acknowledgments

This work has been partially supported by the EC-funded project CoSyne (FP7-ICT-4-24853).

4. REFERENCES

[1] L.P. Huberdeau, S. Paquet, A .Désilets 2008. The Cross-Lingual Wiki Engibe: Enabling Collaboration Across Language Barriers *Proc. of WikiSym 2008.*

[2] A. Kumaran, N. Datha, B. Ashok, K. Saravanan, A. Ande, A. Sharma, S. Vedantham, V. Natampally, V. Dendi and S. Maurice 2010. WikiBABEL: A System for Multilingual Wikipedia Content. *Proc. of AMTA Workshop on Collaborative Translation: Technology, Crowdsourcing and the translator perspective.*

[3] Y. Mehdad, M. Negri, and M. Federico. 2010. Towards Cross-Lingual Textual Entailment. *Proc. of NAACL-HLT 2010.*

[4] Y. Mehdad, M. Negri, and M. Federico. 2011. Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment. *Proc. of ACL-HLT 2011.*

[5] M. Negri and Y. Mehdad. 2010. Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush. *Proc. of NAACL 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk.*

[6] M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti 2010. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. *Proc. of EMNLP 2011.*

[7] V. Nastase, M. Strube, B. Börschinger, C. Zirn, and A. Elghafari. 2010. WikiNet: A very large scale multi-lingual concept network. In *Proc. of LREC 2010.*