A model of translation management systems for multilingual documents

M.H.Traicu and G.Prostean
Politehnica University of Timisoara, Timisoara, Romania

Abstract - This paper suggests a model for translation management systems meant to overcome the problems faced by companies at a global level. Large international organizations do not have to overcome only the problem created by large amounts of data but also by multilingual documents used across the organization. Until now the research related to translation management systems was focused on in-text matching content and re-use capabilities but does not provide solutions for automatic translation. This paper proposes a model for multilingual document management, providing an algorithm for automatic translation management based on the DITA architecture of content components and machine translation algorithm.

Keywords: translation management system, computer environment, translation algorithm

I. INTRODUCTION

A. Background

The problem of managing documents from inception to publishing is a challenging problem faced by any industrial organization. The inability to find previous versions of a document or the incapacity to satisfy publishing requirements are only two of the effects of poor document management systems. Furthermore, the necessity to manage translations for multinational companies with businesses in different countries has increased in terms of cost and technology.

Document management systems have the ability to solve these problems by offering a single repository for content, a shared creation environment for the authors and multiple publishing methods to respond to all the requirements defined by the used of the system.

Despite the interest in the topic of information systems for document management, the research topic is still young and the literature review does not even provide a unified definition of content management, only a common view among the elements that have to be included in such a system: a common repository steered by a software application that provides a shared authoring environment.

The translation of components in different languages is a critical feature of a document management system DMS. A modern multi-language DMS has the ability of “in context match” which means that when a content component is modified the system automatically knows the secondary languages that the component needs to be translated into.

In terms of functionality the TMS has to determine for a specific document and target language the source components without translation. These components are exported from the “pool of assets” to the translation management system. Within this topic, the issue of availability of the content of the translation memory must be addressed. In this sense, the content has to be kept in a standard format so that it can be transferred without difficulty. The common format for multi-lingual content is XML-UNICODE format.

Until now, the research for translation management systems was focused on dedicated documents, for example legal documents or pharmaceutical documents. These documents are characterized by strict formatting and are usually created using templates. The translation management systems for these documents provide in-text matching content and re-use capabilities but do not have automatic translation capabilities.

This paper tries to provide a solution for automatic translation of different content modules and limit the necessity of user intervention, thus eliminating unnecessary effort associated with manual translation.

B. Research topic and objectives

The topic of translation management systems is an ample topic that involves strategies, tools for creating and translating documents, data models and algorithms for automated translation.

Even though there are many research papers that present translation algorithms for specific types of documents, there is a gap in the research for automatic translation algorithms based in systems created on content modules.

In an attempt to bring new insights to this research topic, this paper will focus on the following research question:
What algorithm applies for automatic translation management systems based on content modules?

C. Paper structure

The paper is organized into 5 sections. Section 1 presents the background of the research topic and the research question.

Section 2 provides a synthetic theoretical background where the concepts related to translation management systems are presented.

Section 3 presents the research and analysis methods used to solve the research question.

Section 4 provides the algorithm model for automatic translation management systems. The last section, number 5, presents the conclusions of the paper.

II. LITERATURE ANALYSIS

According to Lakritz, a translation management system (TMS) is a “computer environment which automatically detects when a document, data stream, or non-text file in a master language has been updated and notifies the user which corresponding documents, data streams, or non-text files in the other languages require translation.” [1]

Literature review suggests that the most recent contributions in the field of TMS are based on the DITA architecture. DITA is an open-source “XML-based architecture used for authoring, producing and delivering information”. Besides providing a creation and sharing environment DITA also provides a translation management tool called localization based on module components of documents. [2]


Doyle (2008) suggests that “technical communicators” need “to reinvent themselves and develop the topic of translation management systems based on the DITA technology.”[4]

According to the same author, “the key behind Darwin Information Typing Architecture is to create content in small chunks called topics”. Doyle explains that “these topics are assembled

The disadvantage of DITA architecture is “the translation of content is done by translation centers or individual translators” and not automatically using a translation machine. [2]

Lakritz emphasizes on the fact that current translation applications “use approaches that make it difficult to manage the translation and synchronization of multilingual content.”[1]

III. RESEARCH METHODOLOGY

The research methods used for writing this paper is based on literature synthesis and rational analysis. The comprehensive study of literature sources is the basis of the research process.

The literature review is the base for analyzing the translation management systems and for understanding the topic.

Also, the solution to the research questions is constructed based on the literature analysis.

IV. TRANSLATION MANAGEMENT MODEL

A. Model structure

The model is based on modular content components. The basic philosophy of the document management system is the use of common paragraphs or document components.

The document components are stored in a common database or a pool of assets. This is a double data base both for components and metadata. The paragraphs can be re-used in different documents based on pre-defined rules. Once a reused component is updated in one document it is automatically updated in all documents re-using that component [6]

The source component in the master language can be a paragraph, a picture or a chart. When this component becomes part of a document, the component can be reused.

Figure 1 shows that when a component is used in a document a bidirectional link between the repository and documents is created. Therefore, when searching by document name, an user receives a list of all the components that are used in a document. On the other hand, is the search function is done by component the
user can see a list of all the documents containing that component. This bidirectional link most commonly represents a HTML document structure. For example, document x which is formed by the 2 modules: module A and module B has to be stored in Romanian (master language) and English secondary language. The HTML structure for this document would be:

Document x/
  Romanian/
    Index.html
    Module.html
  English
    Index.html
    Module.html

The most important advantage of using a bidirectional database is the fact that when a component is modified in one document, it automatically is updated in all the documents where it is used. Furthermore, the latest version of the component is saved in the repository. Previous versions of the component are also archived in the repository and can be accessed using search by use date.

Maintaining a database for translated modules provides a stable and secure way of managing large amounts of corporate documents. [5]

B. Algorithm for automated translation

The algorithm for automated translation is based on the phrase-based statistical translation system implemented by Farzindar[4] improved with the DITA content selection method. The flow chart of the algorithm is presented in figure 2.

The first phase of the algorithm consists of identifying modules of a document that have been updated in the master language, or any new modules. The modified components are identified using the DITA open source system.

The second step of the algorithm is translating the differences into the secondary languages. The proposed algorithm uses different translation models to select the best option for the translation.

After identifying the differences, the sentences containing changes are translated using the following translation engines: GIZA ++ [7], Phramer[8] Pharaoh [9] and Ramses [10]. The results are compared using a text editor with reviewing capabilities, such as word. The translated version of the content is selected by counting the repetitions for each word.

The last part of the algorithm consists of automatically inserting the translated content in the XML files and back into the pool of assets.

Figure 2. Flow chart of the translation algorithm

Figure 3 presents the whole architecture of the document management system with the translation system automatically performed based on the algorithm defined in this section.

In terms of functionality the TMS has to determine for a specific document and target language the source components without translation. These components are exported from the “pool of assets” to the translation management system. Within this topic, the issue of availability of the content of the translation memory must be addressed. In this sense, the content has to be kept in a standard format so that it can be transferred without difficulty. The common format for multi-lingual content is XML-UNICODE format.
One of the requirements for the TMS is to identify the components that don’t have translation and translate only those. After finishing the translation of the components, they are transferred back to the CMS where they are stored and linked to the appropriate source component.

V. CONCLUSIONS

The purpose of this paper is to provide a theoretical algorithm for a translation management system that provides both in-text matching content and automatic translation capabilities.

The model described for translation management systems is based on DITA and a statistical translation machine engine.

The xml format used in this model makes it an easily implementable and scalable translation management system.

The innovation of the paper consists in using the modular content components and a phrase-based machine statistical translation system to automatically update translation in a multilingual document management system.

Due to time constraints the study is limited to theoretical concept. The study was conducted over a period of three month. Therefore, future research will focus on implementing the algorithm and testing it in different organizations with different document flows in order to validate it.

REFERENCES