On Localization of Enterprise Information Systems

Goutam Kumar Saha

CDAC, Salt Lake, Sector-V, Kolkata 700091, India
sahagk@gmail.com

Abstract. This paper describes how to localize various output information, including alert messages of an enterprise information system by the XML based Computational Linguistics Markup (CLM). To display a web document or program message or to display a user interface in an appropriate, locale and culture specific translated form, is an important and complex task in the I18N & L10N or Globalization process. In L10N we need to address various locale and cultural aspects, for examples, Naming, formats of date and time, number, icons, symbols and colors etc including legal aspects for proper customization of a product. Language localization denotes the process of translating a product into different languages. Software localization addresses the messages that a program of an Enterprise Information System (EIS) presents to a user need to be translated into various languages. This is very important in the Internationalization & Localization process for addressing language and locale specific various must-do issues as an aid to an easier faster and more meaningful translation process for an EIS web content and answers of the web applications in an EIS. The approach presented here relies on a 3-Layered XML Schematic scheme. By using the CLM, while internationalizing a product, we can do localization easily even without having much linguistic resources on a source human language. The work of this paper is a significant step forward toward globalizing the information system of an enterprise at lower cost for higher gain. A next generation EIS would provide such challenging features of dynamically localized presentation layer with this CLM.

Keywords: *Enterprise information systems (EIS), Enterprise language, Web-based logistics, XML and XML schema, Internationalization and localization of Web content, Computational linguistic markup, ontology, Interoperability models, Software architecture*

1. INTRODUCTION

Today, in the era of global competition, localization of the information system [1-3] of an enterprise is must enabling easy reach to its customers at every part of the globe irrespective of its language and culture. The term "Internationalization" refers to the process involved in the design and development of a product, application or document-content that enables easy localization for target-audiences that vary in culture, region, or language. "Localization" is the process on adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific target market or a "locale". Localization is abbreviated as "L10N", after the number of letters between 'L' and 'N'. We can view internationalization (I18N) as

a process that enables L10N. To display a web document or program message or to display a user interface in an appropriate, locale and culture specific translated form is an important and complex task in the I18N & L10N or Globalization process. In L10N we need to address various locale and cultural aspects, for examples, Naming, formats of date and time, number, icons, symbols and colors etc including legal aspects for proper customization of a product. Language localization denotes the process of translating a product into different languages. Software localization addresses the messages that a program of an Enterprise Information System (EIS) presents to a user need to be translated into various languages. This paper describes how to localize various output information of an enterprise information system by the XML based Computational Linguistics Markup (CLM). This is very important in the Internationalization & Localization process for addressing language and locale specific various must-do issues as an aid to an easier faster and more meaningful translation process for an EIS web content and answers of the web applications in an EIS. The approach talked about here relies on a 3-Layered XML Schematic scheme. On using the CLM, while internationalizing a product, we can do localization easily even without having much linguistic resources on a source human language. The work of this paper is a significant step forward toward globalizing the information system of an enterprise at lower cost for higher gain. The next generation EIS would provide user interactions at users' local languages dynamically apart from an overwhelming functional complexity combined with the requirements of higher interoperability, flexibility, extensibility, adaptive- ness and of course with higher dependability.

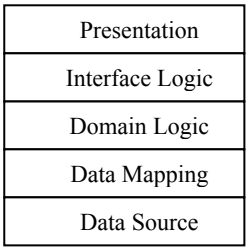

Figure 1. Logical Architecture of Enterprise Information System

Nowadays, logical architecture of an EIS in [2] has five logical layers (as shown in figure 1) viz., *Presentation, Interface Logic* for User Interface, *Domain Logic, Data Mapping, and Data Source.* Such model is to add a mediating layer that works among the primary layers (e.g., User Interface, Domain Logic, Data Source) to obtain a higher independence among them and, then, to permit their evolution or substitution autonomously, decreasing considerably their impact in the other parts of the system.

2. THE THREE-LAYERED XML – BASED CLM

The proposed 3-Layered XML Schema [4-5] aids to markup for both the syntactic as in [6] and semantic metadata information in the structure of an XML document. This approach is a low-cost solution to I18N and L10N [2,7-10] for producing meaningful translation as the various formative, semantic, grammatical and locale specific embedded information of EIS are found to be very important to both the internationalization and localization processes. An XML schema is any type of model document that defines the structure of an XML document. We can create XML schemas using basic XML. In human languages, we often find that a word has several meanings (i.e., word sense ambiguity) at various content contexts (or content domain of a paragraph of a web page). Similarly, a word may have several linguistic parts of speech (i.e., POS ambiguity). For an example, the word "light" has several POS namely, verb, adjective, noun. Again, a metadata about a sentence helps in parsing during the machine translation of web content. The proposed 3-Tier XML Schema approach uses three schemas for web content. The first schema is meant for content domain, the second schema is for sentence level metadata and the third one is meant for the word level metadata or markups. We need to validate an XML document against the proposed three schemas to examine whether the XML content is well formed to conform to the schemas. This 3-tier schema scheme is also useful for the Translation Memory processes to keep context markups when Internationalization & Localization developers use this scheme for both source and target text. We develop the first XML schema that contains various categories on content domain. The second XML schema contains various categories on sentences. The third XML schema contains various Parts-of-Speech categories on words. The proposed scheme uses three XML elements namely, content domain, sentence category and POS category. The schematic block diagram of the proposed 3-Tier or 3-Layered XML Schema approach is shown in figure 2. Content domain includes various contexts namely, sports, information technology, medicine, travel, personal, mathematics and romance etc. Sentence categories include simple, compound, complex, proverbial, taunt, suspicion, active & passive voice, direct and indirect speech etc. Parts-of-Speech categories include noun, pronoun, verb, adjective, adverb, preposition, postposition, interjection, conjunction and indeclinable etc. A content author for EIS having school level grammatical knowledge will not find any difficulty on using such markups because this scheme does not limit one to add an appropriate markup as an attribute. Content author may not use such three level markups at all parts (not for all words and sentences) of a document. Markups need to be used only at the sensitive or difficult parts or ambiguous parts of a document. For some languages, a content author even may not need to add finer sub-category markups at his /her document. Metadata information about the domain, sentence type or specific words will help translators to do better quality work or to do the work quickly**.** If translators know that a word belongs to a specific domain then they can go to a terminology database and check the word; thus, even for human translators this 3-Tier or 3-layer schema will be helpful. One cannot do an accurate translation without such information. The proposed 3-tiers model for dealing with various information on Metadata annotation, sentences and linguistic unit annotations, is illustrated in example 1. The result of

such segmentation is an annotated document, which is very useful to auto-construct Translation Memory and linguistic resources. The first level schema on content domain helps in finding appropriate terminology for the document words. The translation parser gets benefited from the syntactic or formative markup in the second level schema meant for sentence level and the semantic information helps in getting more meaning translation for a given sentence in the content written in a source human language. The third level schema is to embed word level syntactic and semantic metadata information for a word in a sentence. Importance of the 3-layered schema is to find solution for both word sense disambiguation and POS-level disambiguation. For example, in English: the word "bat" has multiple meanings- (a) "a bird" (in the content-domain of zoology/animal) or (b) in the content domain of sports it means "a playing instrument" to hit a ball (like cricket bat). In both cases, parts-of-speech (POS) of "bat" is noun (finer category- common noun) only. In many languages, there are many common words that have multiple meanings (word sense ambiguity) at various contexts (though POS category may remain same). Only based on the context of content domain and sentence, we understand the appropriate meaning of such words having word sense ambiguities. For example, the English word "bank" may mean a financial institution or a stretch of rising land at the edge of a stream. The 1st level schema (content domain markup) is useful for marking the context information for a paragraph of translatable content. The 2nd level schema (sentence level markups) takes care of translatable proverbs, idioms, dialect and usage etc for any human language in the world. The 3rd level schema (word level markups) is to obtain the most appropriate meaning of "a word" (having POS ambiguity with multiple POS and word sense ambiguity) in a sentence inside text content. Content author will not find any difficulty on using such markups because this scheme does not limit one to add an appropriate markup as an attribute. Content author may not need to use such three level markups at all parts (not for all words and sentences) of a document. Markups may be used only at the sensitive or difficult parts or ambiguous parts of a document. For some languages, a content author even may not need to add finer sub-category markups at his/her document. Finer POS categories are useful for handling pragmatic (deeper semantic) issues of content.

3. THE CLM FOR INTERNATIONALIZATION & LOCALIZATION OF EIS

CLM approach incorporates morphology, syntax, semantics and pragmatics of human languages, which are very essential for better machine translation. In order to translate one language into another, one has to understand the grammar of both languages, including both morphology (the grammar of word forms) and syntax (the grammar of how words are combined to form sentences). In order to understand syntax, one has to also understand the semantics of the vocabulary, and even to understand something of the pragmatics of how the language was being used. The novelty of the CLM- based document modeling approach is that it helps us to translate one language to another without knowing grammar of the source language.

CLM enables machine translation even if we do not have much linguistic resources of source language. Metadata information about the domain, sentence type or specific words will help translators to do better quality work or to do the work quickly. If translators know that a word belongs to a specific context domain then they can go to a terminology database and check the word, thus, even for human translators this 3-layered schema will be helpful. One cannot do an accurate translation without such information. XML Schema authors should prefer to use attributes for metadata information because of their better flexibility and portability. XML element tag is in between "<" and " >". XML element tag is used to hold metadata on author's text that helps a machine translator in translation process. Author's text (for translation) is in between "<>" and "</>". Remark or comment is in between "<! –" and " -->". Examples have been added as tutorial so that readers can apply this 3-tier or 3-layered scheme at their work. We may consider an author text or web content that has a sentence, for example, "She played in Shakespeare." CLM Markup as shown in example-1 is meant for word - sense disambiguation.

Example-1 CLM Based EIS Information Annotation

```
<content_domain name="literature" type="drama">
.... <sentence_category name="semantic" type="demonstrative">
She <pos_category name="verb" meaning="acted"> played </pos_category> in
<pos_category name=noun type="proper" meaning="a title of a Drama">Shakespeare
</pos_category>.
<!-- here, "played" implies the verb "acted" -->
</sentence_category>
..... </content_domain>
```

Example-2. An example on Markup for Javascript ToolTips Text

```
<sentence_cat name="scripttitle_value">
<!-- Markup for Javascript ToolTips text on events like ONMOUSEOVER -->
<A HREF="/tips/page2.asp"      ONMOUSEOVER="this._tip='We <FONT COLOR=red>
                    <B>simplify</B></FONT>
                   DHTML'">        DHTML </A> </sentence_cat>
```

Example-3. Markup for ToolTip in EIS

```
<!-- Word-Level Markup for Tool Tip text word embedded inside an Image -->
 <para> Click here
<image source="begin.jpg" alt="begin" />
<pos_cat name="alt_value"> begin </pos_cat>
 to play now.
</para>
```

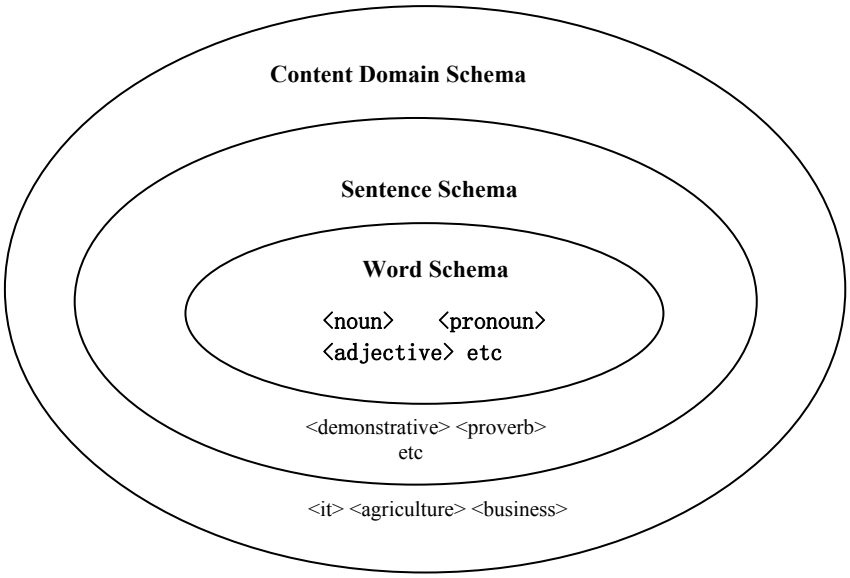

Figure 2. Computational Linguistic Markup Based on 3-layered XML Schema

Example-4. I18n/Localization sensitive Markups used in Postal Address

```
<pos_cat name="noun" meaning="Post Office Box">P.O. Box</pos_cat>4123
<!-- P.O. Box or Postfach (Germany) Or Case Postale (France) in Mailing Address -->
<pos_cat name="noun" meaning="Postal Index Number or ZIP">PIN</pos_cat>700059
<!-- PIN (India) or ZIP (USA) in Mailing Address -->
<pos_cat name="noun" type=”common” meaning="Village">Gram</pos_cat> Dignagar
<!-- here, Gram indicates a village or county (not the measurement unit) -->
<!-- Markup example using Road and lane -->
 Netaji Subhas Road,
Shastribagan<pos_cat name="noun" type=”common” meaning="lane">Goli</pos_cat>
<!-- here, Goli indicates a narrow road or a lane -->
```

4. CONCLUSIONS

The computational linguistic markup based on three- layered XML schema approach is of an immense help in the process of Internationalization and Localization of EIS at the presentation layer. If content author follows CLM based

Internationalization, then Localization to any target language becomes easier. The work aims to ease Localization to any target human language without having knowledge in the source human language at the cost of extra metadata added while internationalizing a web content or software. XML code here is tested as well formed and is tested on web browsers. CLM – based XML document is also validated against the schemas. The paper provides a sound and functional schema, which takes into account the content domain, sentence and word. All the techniques, as described in this paper, relate to how it can be applied from the computation linguistics domain. In addition, with the word, it could be a multi-modal schema, incorporating sound bites of pronunciation. In future, we may describe word meaning in some other standard way, perhaps to hook it up with UNL or WordNet (relations between synonym sets) sense numbers which might further help in automating translation process.

REFERENCES

1. M. Fowler, *Patterns of Enterprise Applications Architecture,* (2007). http://martinfowler.com/isa/index.html (Accessed May 7, 2007).
2. M. Daniel, *The Architecture of Enterprise Information Systems,* (2006). http://moisesdaniel.com/wri/eisa.pdf (Accessed May 3, 2007).
3. Y. Yusuf, A. Gunasekaran, and M.S. Abthorpe, Enterprise information systems: a case study of ERP in rolls-royce*, International Journal of Production Economics.* Volume 87, pp.251-266, (2004).
4. G.K. Saha, *Computational Linguistic Markup (CLM),* W3C Archive, (2005). http://esw.w3.org/topic/its0908LinguisticMarkup (Accessed May 2007).
5. G. K. Saha, A novel 3-tier xml schematic approach for web page translation, *ACM Ubiquity.* Volume 6, Number 43, pp.1-16, (2005).
6. R.P. Sinha, *Current English Grammar and Usage* (Oxford University Press: 2003).
7. *W3C Archive* (2006). http://www.w3.org/TR/its (Accessed May 5, 2007).
8. *W3C Archive* (2006). http://www.w3.org/TR/2006/WD-its-20060222 (Accessed May 5, 2007).
9. R. Ishida and S. K Miller, *Localization vs. Internationalization,* W3C Archive (2006). http://www.w3.org/International/questions/qa-i18n (Accessed May 4, 2007).
10. F. Sasaki, From characters tow web services to internationalization is everywhere, *ACM Ubiquity.* Volume 6, Number 47, pp1-6, (2005).
11. *W3C Archive* (2006). http://www.w3.org/International/quicktips (Accessed May 6, 2007).
12. G.K. Saha, English to bangla translator – the banganubad, *International Journal of Computer Processing of Oriental Languages*. Volume 18, Number 4, pp.281-290, (2005).