IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 5 (Jul. - Aug. 2013), PP 14-18
www.iosrjournals.org

A hybrid English to Malayalam machine translator for Malayalam content creation in Wikis

Nithya B[1], Shibily Joseph[2]

[1]*(Department of Computer Science and Engineering, Government Engineering College, Thrissur, India)*
[2]*(Department of Computer Science and Engineering, Government Engineering College, Thrissur, India)*

Abstract: *Wikis are web-based software platforms that allow users to collaboratively create and edit web page content, through a Web browser. The ease-of-use and open philosophy of wikis have made them a simple and quick method for collaborative knowledge management. As wikis mostly contain natural language, the integration of wiki systems with automated Natural Language Processing (NLP) techniques can be considered to support users in information analysis and content development. The presence of the Dravidian language, Malayalam, is comparatively low in popular wikis like Wikipedia. As a solution to this problem, this paper considers the idea of integration of a Hybrid Machine Translator service, a combination of Statistical Machine Translator (SMT) and Translation Memory(TM), with wiki system for automated content generation in Malayalam. A system architecture providing this integration is presented. The whole process is supposed to speed up the content creation in wikis and bridging the language barrier which restricts knowledge dissemination in the web.*

Keywords *- Content creation, Hybrid Machine Translation, Natural Language Processing, Statistical Machine Translation, Translation Memory, Wiki*

I. INTRODUCTION

A wiki is a website which allows its users to collaboratively add, modify, or delete content via a web browser. Nowadays they have become a powerful tool for content and knowledge management in the web and are being used by simple online communities to very big organizations. The ease of content creation in wiki has attracted a lot of contributors from round the globe. The widespread popularity of wikis is attributed to its open-editing policy as anyone can easily become a content contributor.

The word Wiki is actually the short form for WikiWikiWeb which is derived from the Hawaiian word ‘wikiwiki’ meaning fast or quick. The very first wiki was developed by Ward Cunningham in 1995. Wikipedia, Wiktionary etc. are some very popular examples for wiki. Wikis are powered by underlying wiki software. A wide range of wiki systems are available like MediaWiki, TWiki, XWiki etc.

A wiki system is based on the following design principles [3]. A wiki page can be edited via a simple browser interface. The syntax for creating content is quite simple and even a lay man can easily take part in content creation process. Hyperlinks can be easily added to the wiki structure. There exists versioning support for wiki edits and so it is easy to roll back to the previous version of a page. Most of the wiki systems provide an unrestricted access to users. But some employ access restriction in the form of username and password. Wikis are an ideal tool for collaborative editing. As soon as someone creates content, others can contribute to it, extend it, correct it, etc. Search function exists in wikis which makes it easy to find the required content in wikis. All these features have made wikis one of the most relied upon and referenced information sources in the web.

Natural Language Processing (NLP) is a branch of computer science that employs various Artificial Intelligence (AI) techniques to process content written in natural language. NLP has now become an important tool for information and knowledge management. Wiki content is mostly written in unstructured natural language. Also wikis do not enforce the users to structure pages and hence they often end up as a collection of unmanageable pages. Human ability to retrieve or organize content becomes limited due to the large size of wiki content owing to its open editing policy. Hence automated NLP techniques can be applied in the case of wikis to provide an intelligent support for the wiki user.

The traditional method of content creation in wikis is to provide the content manually from scratch. This method is highly suited for those languages like English with a large number of content contributors. But the last few decades saw a great rise in the number of Internet users ranging from highly educated people to school children and even people who are not that good in English has also started making use of Internet. This has led to a new situation where there is an urge for providing web information repositories in local languages too rather than just in English. Popular wikis like Wikipedia suffers from an uneven distribution of articles i.e., majority of the articles are available in English language only. Malayalam is a traditional Dravidian language with very low presence in the web and wikis. The approach of manual content creation is not practical in languages like Malayalam as the number of content contributors is quite low. So the automation of the content creation process will be of great use.

Machine translation is one of the mostly researched tasks in NLP. It is the process of translating text from one natural language into another natural language using computers, with or without human assistance. As the web content is mainly available in English language, a machine translator can be made use of to make the vast amount of information in the Internet available in languages other than English. In the case of wikis, the translator can be used for automated content creation in languages other than English which boosts the localization efforts in the target language.

The rest of the paper is organized as follows. Section 2 describes the works related to wiki-NLP integration and English to Malayalam translators. Section 3 discusses the traditional translation approaches used and section 4 describes the Hybrid Machine Translator for English to Malayalam translation. Section 5 gives the design of the architecture for the integration of the machine translator with wiki. Section 6 concludes the paper.

II. RELATED WORKS

There exist several approaches aimed at automating tasks on behalf of Wiki users. In the field of automatic and semi-automatic Wiki content editing, a wide array of bot frameworks existed to perform tasks like spell checking, line checking etc. These techniques helped in improving the quality and maintenance of wikis but do not perform any natural language processing on the wiki content. They functioned on the technical and syntactic level and not on the semantic level.

Semantic extensions to Wiki systems, based on Semantic Web technologies like Web Ontology Language (OWL) ontologies and Resource Description Framework (RDF), provide the means for further structuring and automated processing of Wiki content [3]. The focus of this semantic wiki is on semantically annotating the content in the wiki, which in turn allows structured search. Semantic wikis make wikis understandable by machines, by adding attributes and properties to the content as categories and annotated hyperlinks. The drawback is that they rely on explicitly provided semantic information that has to be manually added by the user. Semantic MediaWiki and IkeWiki are some popular semantic wiki implementations.

The idea of integration of NLP techniques with wikis was first proposed by Witte and Gitzinger[4]. They proposed a multi-tier architecture to connect wikis to services providing NLP functionality, which are based on the General Architecture for Text Engineering (GATE). The services provided are automatic summarization of wiki content, question answering, and index generation to facilitate the search process. The system uses a separate application to apply the services to the wiki content. The application reads the content of the wiki page, applies NLP algorithms to it, and writes the modified page back to the wiki. Later Johannes Hoart, Torsten Zesch, Irynaurevych [5][6] proposed an enhanced architecture, Wikulu, which provided an intelligent user interface augmented with NLP services. This architecture allows NLP techniques to be integrated into existing wiki platforms. It integrates a wide range of NLP algorithms such as keyphrase extraction, link discovery, text segmentation, summarization, text similarity etc. for better content management.

Regarding the developments in the field of English to Malayalam translator, there are only a few efforts. A model for rule based English to Malayalam translation was proposed by Remya Rajan, Remya Sivan, Remya Ravindran, K.P Soman[7]. The development of a Statistical Machine Translation (SMT) system for English to South Dravidian languages like Malayalam and Kannada by incorporating syntactic and morphological information was proposed by Unnikrishnan P, Antony P J and Dr. Soman K P [1]. A methodology for translating text from English to Malayalam using statistical models was proposed by Mary Priya Sebastian, Sheena Kurian K and G Santhosh Kumar [8]. This method incorporates morphological information into the parallel corpus with the help of a parts of speech tagger. AnglaMalayalam, a rule based and interlingua based English to Malayalam translator, a part of AnglaMT, was developed by CDAC and IIIT jointly . Integration of Translation Memory(TM) with AnglaMT, a RBMT, was proposed in [9].

III. TRADITIONAL TRANSLATION PRACTICES

For content translation of wiki sites, there exist no processes and tools. But such processes and tools exist and are being used for translating content in traditional industrial environments. Three main techniques used for translation in the industrial environment [10] are sequential translation, parallel authoring and incremental just-in-time translation.

Sequential translation is popular with very large industrial organizations such as automobile and computer manufacturers, who need to publish large amounts of product documentation in several languages. In this technique actual authoring is done first and later translation is done by an external translation agency or translation companies. This technique is optimized for translating a document which will not change in the future.

In parallel authoring approach, the documents in all languages are created in parallel by language experts and technical experts. This creates all documents almost by the same time and reduces the inherent delay in translation.

With the incremental just-in-time translation process, changes to web pages are tracked in real-time, and requests for translations are issued as soon as such a change occurs. This technique combines the best of earlier two approaches and is suitable in cases where the document is subject to changes.

None of these translation practices are suitable for the wiki world as it is a collaborative platform for content creation without any centralized control. So some other translation techniques need to be devised for wikis

IV. HYBRID APPROACH TO MACHINE TRANSLATION

To ensure a high-quality product, diagrams and lettering MUST be either computer-drafted or drawn using India ink. The hybrid approach to Machine Translation is a combination of one or more translation approaches. Here the idea is to integrate TM with the traditional SMT for performing English to Malayalam translation.

SMT [2] is a corpus-based machine translation approach in which machine learning techniques are applied to a bilingual corpus to produce a translation system automatically. For an SMT system, a parallel corpus consisting of source and target language sentences and a monolingual corpus consisting of target language sentences is required. The statistical model learns the translation parameters from the corpus and performs the translation.

Every sentence in the target language is considered as the translation of a source language sentence with some probability. The best translation is the sentence that has the highest probability. SMT models views the machine translation as a noisy channel model. If we want to translate a sentence f in the source language F to a sentence e in the target language E, the noisy channel model describes the situation in the following way: Suppose that the sentence f to be translated was initially conceived in language E as some sentence e. During communication e was corrupted by the channel and became f. Now assume that each sentence in E is a translation of f with some probability and the sentence we choose as the translation is the one that has the highest probability, i.e., the sentence e that maximizes the probability $P(e|f)$.

The SMT can be split into two main phases - Training phase and Translation phase. The first phase is the training phase, in which a statistical model of translation is built, using a corpus of texts in both the source and target language. The training phase is itself split into three parts: (i) document collection, where we assemble the corpus of texts (ii) building the language model for the target language from the monolingual corpus i.e., $P(e)$ and (iii) building the translation model from the target language to the source language i.e., $P(f|e)$. The second phase is the translation phase or the decoding phase, which uses a heuristic search procedure to find a good translation of a text.

TM is a mechanism for caching the recently performed translations to aid human translators. It can be considered as a database that stores the source text and its corresponding translation in language pairs called translation units.

V. PROPOSED ARCHITECTURE

The integration of wiki system with a machine translator can be considered for content creation in Malayalam language from corresponding English language content. The architecture for this integration makes the following assumptions:

- English is the master language.
- The source content is available in the master language.
- The machine translator model is statistical coupled with TM.
- The translation process is a Human Aided Translation and requires human intervention to guide the translation process and that person should be fluent in both the source language and the target language.

The design integrates the machine translator service with wiki. A content contributor who wishes to provide content in Malayalam has to first search whether the English language article exists for the corresponding topic. If English content is present then the translator service can be made use of. The service is invoked via a button interface in the wiki page and the user can select the target language. The English wiki page is retrieved then. The sentences in the web page are retrieved one by one and fed to the machine translator system. The < /html> tag identifies the end of the page. The translator system then performs the translation and returns the translated sentences back to the wiki system. The insertion of translated sentences are done in the new wiki page of the target language.

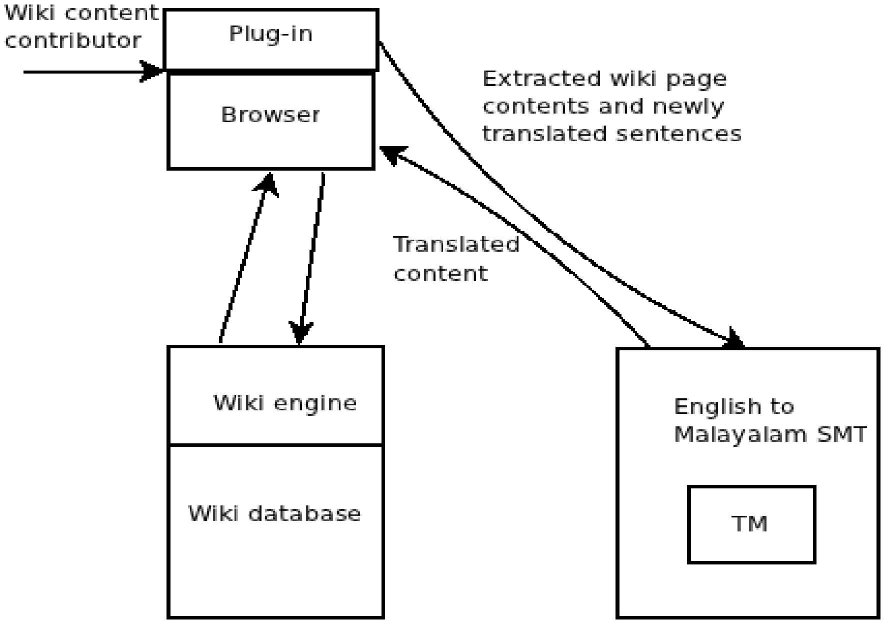

Figure 1: Proposed architecture

The proposed architecture is given in Fig 1. There is a wiki content contributor who interacts with the browser plug-in to invoke the MT service for the wiki. The wiki system is comprised of the wiki engine and the wiki database. The translation is performed by the SMT system.

A statistical model of machine translator coupled with translation memory can be used for wiki translation. Inclusion of post-editing feature allows the human expert to make modifications. As machine translators may not always produce perfect translations, the human intervention enables to correct the mistakes if any. The post-edited translations can be fed to the translator system and can be added to the translation corpus. Thus with each translation, the translator becomes more perfect as the quality of a statistical machine translator output is proportional to the amount and quality of training data available.

VI. Conclusion

The integration of machine translator service with wiki is quite helpful in speeding up the content creation process in wikis in languages like Malayalam. But the effectiveness of this approach is directly dependent on the quality of machine translator output. The time required for translation is also crucial. Since the statistical model of machine translator coupled with TM is used, the performance can be improved by providing quality training data. SMTs require a lot of computing power. More the computing power, better will be the translation. So the SMT can be implemented in parallel on a large cluster of machines to provide better translation.

Acknowledgements

The authors wish to thank the anonymous reviewers for their useful suggestions that helped in improving the quality of this paper.

References

Journal Papers:

[1] Unnikrishnan P, Antony P J and Dr. Soman K P, A Novel Approach for English to South Dravidian Language Statistical Machine Translation System, International Journal on Computer Science and Engineering, 2010.

Books:

[2] Jurafsky, Martin , *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* (Prentice Hall PTR, Upper Saddle River, NJ, USA 2000).

Proceedings Papers:

[3] Sebastian Schaffert, Ikewiki: A semantic wiki for collaborative knowledge management, *Proc. of WETICE, IEEE Computer Society, 2006.*

[4] R. Witte and T. Gitzinger, Connecting wikis and natural language processing systems, Proc. of the Intl. Symposium on Wikis, 2007.

[5] J. Hoart, T. Zesch, and I. Gurevych, An architecture to support intelligent user interfaces for wikis by means of natural language processing, Proc. of the International Symposium on Wikis, ACM, 2009.

[6] D. Bar, N. Erbs, T. Zesch, and I. Gurevych, Wikulu: An extensible architecture for integrating natural language processing techniques with wikis, in Proc. of the ACLHLT System Demonstrations, ACL, 2011.

[7] Remya R, Remya S, Remya R, K.P Soman, Rule Based Machine Translation from English to Malayalam, International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009.

[8] Mary Priya Sebastian, Sheena Kurian K and G. Santhosh Kumar, English to Malayalam Translation: A Statistical Approach, Amrita ACM-WiC, India, 2010.

[9] Nishtha Jaiswal, Renu Balyan and Anuradha Sharma, A step towards Human Machine unification using Translation Memory and Machine Translation System, International Conference on Languages, Literature and Linguistics, 2011.

[10] Alain Dsilets, Lucas Gonzalez, Sbastien Paquet, Marta Stojanovic, Translation the Wiki Way,in Proc. of the International Symposium on Wikis, ACM, 2009.