A MULTI-AGENT SOLUTION FOR MANAGING COMPLEXITY IN ENGLISH TO SINHALA MACHINE TRANSLATION

B. HETTIGE, A. S. KARUNANANDA & G. RZEVSKI
Department of Computational Mathematics, University of Moratuwa, Sri Lanka.

ABSTRACT

Machine translation turns out to be an inherently complex process requiring serious attention to morphological, syntactic and semantic complexity within both the source and the target languages. Most of the existing approaches to machine translation (MT) circumvent the complexity with the assumption that morphological, syntactic and semantic analysis can be done independently and sequentially. This has resulted in depriving us of the opportunity to use the language complexity to generate high-quality translations. In view of this, research has been conducted to develop a multi-agent systems solution for MT that uses the language complexity as an opportunity for generating a more realistic translation from English to Sinhala. This multi-agent solution primarily comprises a six-agent swarm to deliberate on morphological, syntactic and semantic concerns of the source and the target languages without being constrained to operate in a sequential manner. These agents use the ontology of corpora and dictionary of two languages. This approach is inspired by the fact that people understand a sentence by incrementally reading through words while simultaneously considering the syntax and semantics. As such, when the system progresses in identification of words one by one, both syntactical and semantic concerns are entertained up to the current point of reading. As a result, initially decided words may be changed due to the present concern of morphology, syntax and semantics. A translation system has been implemented on the multi-agent system development framework named MaSMT. Experiments show that the multi-agent solution for MT gives promising results for translating sentences of an average length and further research has been carried out to accommodate translation of long sentences.

Keywords: machine translation, multi-agent systems, Sinhala language.

1 INTRODUCTION

Machine translation (MT) has been a research concerning the translation of text or speech from one natural language (source language) to another natural (target language) [1]. According to the complexity of the natural languages, MT has become a reality in research since last few decades. At present, a number of approaches are available to MT, including human-assisted, rule-based, statistical, example-based, knowledge-based and hybrid. Using these approaches thousands of MT systems have been developed to translate natural languages around the world. Among these approaches, statistical approach has been the widely used successful approach for the MT since last decade. Google translator [2], Moses [3], SYSTRAN [4] are the successful statistical-based MT systems in the world. In addition to the above, rule-based approach is powerful and commonly used approach for MT especially for target languages that have different writing methods for the written and spoken. Further, rule-based systems are capable to provide grammatically correct translation by using morphological, syntactical and semantic rules of the source and target language. Number of MT systems have been designed through rule-based approach, including Apertium [5], GramTrans [6], Anusaaraka [7] and BEES [8].

In view of the inherent complexity of the natural languages, MT itself turn out to be a complex task requiring analysis and generation of morphological, syntactic and semantic information of both source and target languages, respectively. Furthermore, a successful MT system requires providing suitable solutions to handle translation issues such as word ambiguity, idiomatic phrase translation, different syntax structures, translate anaphors and co-reference. Much of existing approaches to MT circumvent the complexity with the assumption that morphological, syntactic and semantic analysis

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V11-N2-88-96

can be done independently and sequentially. This has resulted in depriving of opportunity to use the language complexity to generate high-quality translations. As a result, at present, even the most successful MT systems in the world could not achieve > 90% accuracy.

In responding to above issue, a research has been conducted to develop multi-agent systems solution for MT that uses the language complexity as an opportunity for generating a more realistic translation from English to Sinhala. This multi-agent system solution primarily comprises six-agent swarms to deliberate on morphological, syntactic and semantic concerns of the source and the target languages without being constrained to operate in a sequential manner. These agents use the ontology of corpora and dictionary of two languages. The multi-agent approach to MT is inspired by the fact that people understand a sentence by incrementally reading through words while simultaneously considering the syntax and semantics. Stated another way, a person never reads all the words in a sentence first, and looks for the grammatical accuracy and the meanings of the sentence at last. Thus, our solution operates in a unique manner where a morphological agent deliberates with syntactic and semantic agents to identify the words incrementally in a sentence. As such, when the system progresses in identification of words one by one, both syntactical and semantic concerns are entertained up to current point of reading. Accordingly, initially decided words may be changed due to the present concern of morphology, syntax and semantics.

Concurrently with the reading through the words in an English sentence, corresponding Sinhala words are also identified from the bilingual database. The initially identified Sinhala words may change when Sinhala syntactic and semantic agents start executing to form a grammatically correct sentence in Sinhala. The translation happens only after reading an English sentence fully.

The system has been implemented with multi-agent development framework MaSMT, which is specially designed to develop MT applications through the multi-agent approach. Translation system consists of six sub-systems, namely English morphological system, English syntactical system, English semantic system, Sinhala morphological system, Sinhala syntactical system and Sinhala semantic system that handle morphology, syntax and semantics of both languages. The English morphological agents are responsible to analyze the English morphology and Sinhala morphological agents are responsible to generate Sinhala word for the given morphological information. English syntax agents are responsible to analyze the syntax structure of the English sentence and Sinhala syntax agents are responsible to generate Sinhala syntax structure for the given syntax information. Semantic agents are also capable of analyzing and identifying the semantic information as required for the translation. Final solution provides through the communication among these sub-systems of agents.

The rest of the paper is organized as follows. Section 2 gives a brief summary of the existing approaches, systems and issues for the MT. Section 3 presents our novel approach to MT. Section 4 explains how system works in practice. Section 5 concludes the paper with a note on further work.

2 MACHINE TRANSLATION APPROACHES, SYSTEMS AND ISSUES

This section briefly describes existing approaches, systems and common issues in MT.

2.1 Approaches to machine translation

A number of approaches are available for MT including direct translation, dictionary-based, human-assisted, statistical, example-based, rule-based and hybrid [1]. By using these approaches, a large number of MT systems are developed for many natural languages.

Direct translation is a primitive approach to MT that replaces the words in source language with words in the target language by using a bilingual dictionary. This approach does not care to do linguistic analysis or processing and uses only a bilingual dictionary for the translation. According to this approach, dictionary lookups may be done with or without morphological analysis or lemmatization. In contrast, dictionary-based systems are commonly used for cross-language retrieval systems and ideally suitable for the translation of long lists of phrases on related language pairs, which has same syntactic and structure. Telugu to English translation [9], Japanese–English dictionary-based MT system [10], Bengali and Hindi to English cross-language text retrieval [11] systems are dictionary-based MT systems.

Human-assisted MT approach is particularly used for Indian families of MT. The human-assisted approach uses human interaction for the pre-editing, post-editing and/or intermediate editing stages [10]. This approach uses human support for the semantic handling in the MT. This type of MT system is also called semi-automated MT systems. Statistical MT approach is by far the most widely studied MT method in the field of natural language processing. This approach tries to generate translations using statistical methods based on bilingual text corpora. Moses [12], Google Translator [2], are the popular Statistical MT Systems.

Example-based MT (EBMT) approach uses bilingual corpus with parallel texts as the knowledge base. These types of MT systems are more suitable for the languages with less resources. Example-based systems are trained through the bilingual parallel corpora, which contain sentence pairs. Kyoto-U [13] and example-based English to Sinhala machine translator [14] have used this approach to MT.

The rule-based approach is yet another approach for MT. This approach gives grammatically correct translation by using a set of rules. Rule-based system consists of morphological, syntax and semantics rules to process the translation. Most of these MT approaches run with in sequence and start with morphological analysis of the source language and end with the target language morphological generation. Anusaaraka [7], BEES [8] are some example for the rule-based translation.

The hybrid MT systems use combine method in two or more approaches especially rule-based and statistical MT approaches. English to Malayalam MT [15] and TransEasy [16] system use this approach to MT.

2.2 Existing English to Sinhala machine translation systems

Google translation is one of the popular free multilingual translation services provided by Google. At present, it supports more than 90 languages, including Sinhala. This translation service offers a web interface, mobile interfaces for Android and iOS, and an API for developers who can use to build browser extensions, applications and other software. [17] Sinhala translation is available on Google since December 2014. According to Google translation, its service limits the number of paragraphs and the range of technical terms that can be translated. However, Google translator helps the reader to understand the general content of a foreign language text and it does not always deliver accurate translations.

BEES is another rule-based MT system, powered by theory of varanageema (conjugation) in Sinhala language [8]. BEES system consists of seven modules namely English morphological analyzer [18], English parser [19], English to Sinhala translator, Sinhala syntax generator, Sinhala morphological generator and Transliteration module [20]. English to Sinhala MT system has been implemented with the use of SWI-Prolog, Java and Prolog Server Pages (PSP).

Liyanapathirana and Weerasinghe have developed a Statistical MT system for English to Sinhala [21]. Silva and Weerasinghe have also developed an example-based English to Sinhala MT [14]. This research has attempted to translate government documents in English to Sinhala through a database of examples.

2.3 Common issues in machine translation

Process of the MT can be primarily realized through morphological, syntactical and semantic level concerns of natural languages. As such, in the outset, an MT system should be able to handle morphological, syntactical and semantic issues correctly to provide successful translation. The first and commonest problem in morphological level has been recognized as the ambiguity in words in a sentence [22]. The meaning of the word is imprecise or open to more than one interpretation or some words have multiple interpretations are the main reasons for the word ambiguity. To avoid the word ambiguity, sufficient context or explanation is required [23]. Translating an idiomatic phrase is another common issue on the MT. An idiom is a common word or phrase with a culturally understood meaning that differs from what its composite words' denotations would suggest [24]. To translate the idiomatic phrases, translation system requires to identify these sets of words as a single unit. Therefore, it is difficult to handle large idiomatic phrases.

A sentence having different syntactical structures is a major issue in syntax level concerns of MT. To handle this syntax-level ambiguity, syntax analysis is required. In addition to that, some structures of the target language differ from the source language (e.g. English has Subject-Object-Verb order and Sinhala comes with Subject-Object-Verb structure. At that point translation is not easy.

Further, a grammatical term for pronoun, which refers back to another word or phrase, known as Anaphora [22] is another issue at the syntax level. This in fact results in semantic-level analysis too. To translate anaphora, MT systems are required to identify the antecedent of the anaphora by matching the features of the anaphor with the nouns of the previous in the nominative form. Then it is required to identify anaphora of the target language. In addition to the above, it is required to translate anaphor from source language to target language; the features of anaphora of source language mapped to the anaphora of the target language. If more than one entry is available for the anaphora in the bilingual lexicon, then the Gender-Number-Person features are matched to get correct anaphora [25].

Co-reference is one of the problems of semantic translation, and it occurs when two or more expressions in a text refer to the same person or thing. To handle co-reference in the translation, complex index is required.

Discourse is another issue of the semantic translation. To handle these semantic issues, it is needed to analyze the discourse structure. In addition to the above facts, some source sentences do not have single correct answers. In the MT point of view, it is an issue to handle through the translation.

3 DESIGN: MULTI-AGENT APPROACH TO MACHINE TRANSLATION

The approach behind the MT is based on the hypothesis that words employ as the building block of natural language understanding and MT process requiring attention to morphological, syntactic and semantic on both source and target languages. This is valid for people who read a sentence word by word or otherwise by locating selected words such as nouns and verbs. Further, consequently, meaning of a sentence is determined by the interaction among words, which draw from all aspects of morphology, syntax and semantics, as appropriate. This solution has been designed as a multi-agent system and implemented through the multi-agent development framework MaSMT [26]. Figure 1 shows the architecture of the MaSMT framework.

The MaSMT framework is especially designed to develop MT application through the multi-agent platform. The MaSMT framework provides two types of agents, namely ordinary agents and manager agents. The manager agent consists of a number of ordinary agents within its control. Further, manager agents can directly communicate with other manager agents (through the message transport agent) and each and every ordinary agent in the swarm is assigned to a particular manager agent. An ordinary agent in a swarm can directly communicate only with the agents in its own swarm

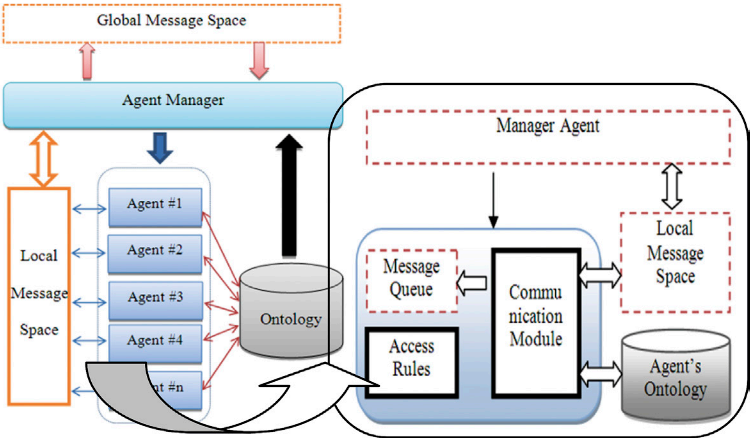

Figure 1: Architecture of the MaSMT framework.

and its manager agent. The MaSMT framework has been implemented by using JAVA and multi-agent system capabilities that have been tested with several applications [27].

English to Sinhala MT system consists of six sub-systems (agent swarms), namely English morphological system, English syntactical system, English semantic system, Sinhala morphological system, Sinhala syntactical system and Sinhala semantic system. In addition to the above, message transport agents and translator agent are the supporting agents. Figure 2 shows the design diagram of English to Sinhala MT system. According to the multi-agent architecture, each subsystem works as an agent swarm. A brief description of each multi-agent swarm is given below.

The English morphological sub-system works as an English morphological analyzer and it is capable to analyze English words and provides the morphological information for the given English word [28]. The morphological system is also a multi-agent system that can communicate with each other agents and provides morphological information as required. The morphological system consists of a manager agent (MaSMT manager) and numbers of morphological agents (agent in the MaSMT) to analyze English morphology. The present system consists of 19 morphological agents to represent 19 English morphological rules.

The English syntax system is the English syntax analyzer of the MT system, which is capable to syntactically analyze the given text. The system also consists of a manager agent and a number of syntactical agents to represent syntax rules. These agents can identify syntactical category of the input words such as noun phrase, verb phrase, subject, verb, etc. After analysis, the system generates the syntax structure of a given sentence of phrase. English semantic system is also capable to analyze the semantics of the given English text. Semantic system gets support from English–Sinhala bilingual dictionary and Internet to collect semantic information as required.

Sinhala morphological system is worked as Sinhala morphological generator and it generates appropriate Sinhala word for the given morphological information. This system also consists of 132 Sinhala morphological agents to support morphological generation through the Sinhala morphological rules [8]. These grammar rules are based on Sinhala word conjugation forms [8]. The Sinhala semantic system also provides semantic information for the Sinhala words that are generated by the system. This system also uses bilingual dictionary and the Internet to get the required information.

Message transport agent is a supporting agent of the translation system that helps to transport messages for each manager. According to the MaSMT architecture each agent in the swarm can

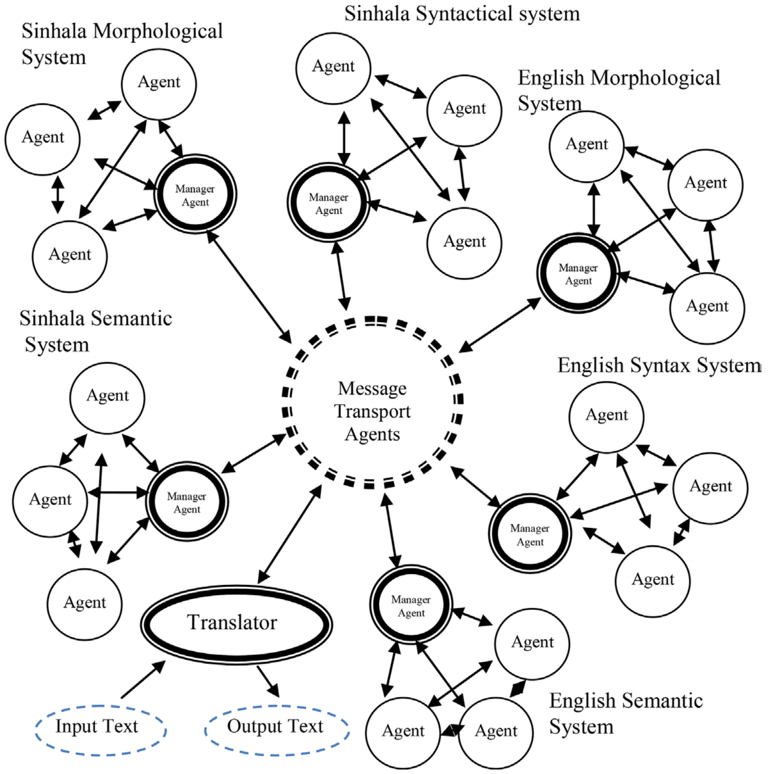

Figure 2: Design of the multi-agent-based English to Sinhala machine translation system.

communicate only for each agent in its own swarm and its manager agent. Managers can communicate only with its agents and other managers. A message passing through the manager agents is done through the message transport agents. The translator is a manager agent that reads input sentence form the graphical user interface (GUI) and provides output for the GUI.

According to the multi-agent architecture, each sub-system is capable to communicate with each other and makes their job as required [29]. The system especially does not give any order to start its task first; as well as some systems may wait until other systems finish their task (syntactical system waits until the morphological system completes its tasks). English to Sinhala MT is done through communication of each of the six sub-systems.

4 HOW THE SYSTEM WORKS

This section briefly describes how MT system works for the given input sentence. According to the translation approach discussed in the paper, translation system works as how a person (non-English person with expertise in target language) translates an English sentence into Sinhala. Most of the professional translators translate a sentence through the two phrases. At first they read all words in the source sentence word by word and get an idea about context and semantics. Then they generate grammatically correct target language sentence with considering the source and target language grammar (especially syntax). This translator uses this mechanism to translate English sentence into Sinhala.

As a first step of the translation, translator agent reads an English sentence from GUI to translate. Then translator reads the sentence word by word on left to right. Assume translator reads a sentence 'good boy reads books.' Then translator gets the first word 'good' and starts the translation process with the support of each sub-system. At this point, all six sub-systems get the word 'good' and start their job. English morphological system analyzes the English word 'good' and sends the morphological information. The English semantic system identifies corresponding Sinhala words for the given English word 'good.' However, English syntax system can start its task after English morphological system finishes its job. Note that, Sinhala morphological system and Sinhala syntax system wait until translator completes the reading of the English sentence.

After that, the translator reads the next word 'boy.' At this point, the English syntax system identifies 'good boy' as a noun phrase and the English semantics system searches for suitable Sinhala words for 'boy' considering the previously selected Sinhala word for good.

Incrementally system reads the word 'reads' as the next word and English syntax system identifies the word 'reads' as a verb phrase. (Note that at this level, English syntax system does the analysis only a phrase level) In this point system consists of a noun phrase 'good boy' and the verb phrase 'reads.' Now English semantic system identifies suitable Sinhala words for the verb 'reads' with considering the previously selected Sinhala words for 'good boy.' As such, system reads all words until the end of the given sentence. According to this approach, this process is same as the person who reads the English sentence and gets some idea about the context. Up to this point, English syntax system does only the phrase-level analysis and semantic system identifies suitable Sinhala words for the each phrase.

When the system progresses in identification of words one by one, both syntactical and semantic concerns are entertained up to current point of reading and initially decided words may be changed due to the present concern of morphology, syntax and semantics.

After reading all the words in the English sentence English morphological sentence identifies the syntactical category of the input English sentence such as subject, verb, object, predicate, etc. Then with communication of English and Sinhala syntactical systems, suitable Sinhala syntax is generated for the given English sentence. Sinhala sentence has a different structure than the English sentence. Therefore, syntactical transform is required to generate the grammatically correct Sinhala sentence. After that, Sinhala morphological system generates the suitable Sinhala word for each Sinhala words considering all the required grammar. Finally GUI system reads the generated target sentence.

5 CONCLUSIONS AND FURTHER WORK

This paper has reported our research on the use of multi-agent system technology to implement English to Sinhala MT. This solution postulated that MT can be done going through words in a sentence by incrementally handling morphological, syntactical and semantics concerns of both source and target languages. This goes beyond the sequential approach to processing of sentences to MT in conventional systems. Our system has been implemented with six sub-systems namely the English morphological system, English syntactical system, English semantic system, Sinhala morphological system, Sinhala syntactical system and Sinhala semantic system. English to Sinhala multi-agent-based MT system is a complex system that handles more than 500 agents, which are interconnected through the six swarms. These sub-systems are not centrally controlled and work as required for the translation. Agents in the system have a degree of autonomy but their behaviour is always subject to certain language-specific rules. This MT system has been implemented on the multi-agent system development framework named as MaSMT. It is evident from the experimental results that our translation system generates promising results when translating sentences with average length and further research are being conducted to accommodate translation of long sentences.

REFERENCES

[1] Hettige, B. & Karunananda A.S., Existing Systems and Approaches for Machine Translation: A Review, *Proc. of the 8th Annual sessions on Sri Lanka Association for Artificial Intelligence (SLAAI)*, pp. 34–40, 2011.

[2] Google Translate Blog, googletranslate.blogspot.com

[3] Moses, www.statmt.org/moses/

[4] SYSTRAN, www.systransoft.com

[5] Forcada M.L., Tyers F.M. & Ramírez-Sánchez G., The Apertium machine translation platform: five years on, *Proc. of 1st Int. Workshop on Free/Open-Source Rule-Based Machine Translation*, pp. 3–10, 2009.

[6] Wiechetek L., Rule-based MT approaches such as Apertium and GramTrans, Online. uit.no/Content/84556/mt.pdf

[7] Chaudhury S., Rao A. & Sharma D.M., Anusaaraka: An expert system based machine translation system, *Proc. of 2010 Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE),* pp. 1–6, 2010.

[8] Hettige B. & Karunananda A.S., A Computational grammar of Sinhala for English-Sinhala machine translation, *Proc. of 2011 Int. Conf. on Advances in ICT for Emerging Regions (ICTer),* pp. 26–31, 2011.

[9] Prasad T.V. & Muthukumaran G.M., Telugu to English Translation using Direct Machine Translation Approach., *International Journal of Science and Engineering Investigations*., **2(12)**, pp. 25–32, 2013.

[10] Okumura A. & Hovy E., Building Japanese-English dictionary based on ontology for machine translation, *Proc. of workshop on Human Language Technology*, pp. 141–146, 1994.

[11] Mandal D., Dandapat S., Gupta M., Banerjee P., & Sarkar S., Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources, *in Working Notes for the CLEF 2007 Workshop*, 2007.

[12] Koehn P. & Hoang H., Birch A., Chris C., Federico M., Nicola B., Brooke C., Wade S., Christine M., Richard Z., Chris D., Ondrej B., Alexandra C., & Evan, Moses: Open Source Toolkit for Statistical Machine Translation, *presented at the Annual Meeting of the Association for Computational Linguistics (ACL),* Prague, Czech Republic, 2007.

[13] Nakazawa T., Yu K., Kawahara D., & Kurohashi S., Example-based machine translation based on deeper NLP, *in IWSLT,* pp. 64–70, 2006.

[14] Silva A.M. & Weerasinghe R., Example Based Machine Translation for English-Sinhala Translations, *Proc. of 9th Int. IT Conference (IITC 2008), Colombo, Sri Lanka*, pp. 27–28, 2008.

[15] Nithya B. & Joseph S., A Hybrid Approach to English to Malayalam Machine Translation., *International Journal of Computer Applications* **8(1)**, pp. 11–15, 2013.

[16] Liu Q. & Yu S., TransEasy: a Chinese-English machine translation system based on hybrid approach*, in Machine Translation and the Information Soup, Springer*, pp. 514–517, 1998.

[17] Google Translate, Wikipedia, the free encyclopaedia.

[18] Hettige B. & Karunananda A.S., A Morphological Analyser to Enable English to Sinhala Machine Translation, *Proc. of Int. Conf. on Information and Automation, ICIA 2006*, pp. 21–26, 2006.

[19] Hettige B. & Karunananda A.S., A Parser for Sinhala Language - First Step Towards English to Sinhala Machine Translation, *Proc. of 1st Int. Conf. on Industrial and Information Systems*, pp. 583–587, 2006.

[20] Hettige & Karunananda A.S., Transliteration system for English to Sinhala machine translation, *Proc. of Int. Conf. on Industrial and Information Systems, ICIIS 2007*, pp. 209–214, 2007.

[21] Liyanapathirana J. & Weerasinghe R., English to Sinhala Machine Translation: Towards Better information access for Sri Lankans, *in Conference on Human Language Technology for Development*, Alexandria, Egypt, pp. 182–186, 2011.

[22] Wikipedia: Ambiguous words, Wikipedia, the free encyclopaedia.

[23] Center T., Ambiguity Reduction for Machine Translation: Human-Computer Collaboration, Online. http://homes.cs.washington.edu

[24] Gaule M. & Josan G.S., Machine Translation of Idioms from English to Hindi, *International Journal of Computational Engineering Research (ijceronline.com)*, **2(6)**, pp. 50–54, 2012.

[25] Suryakanthi T., Prasad S., Prasad T.V., Translation of Pronominal Anaphora from English to Telugu Language, *International Journal of Advanced Computer Science and Applications, (IJACSA)*, **4(4)**, pp. 75–79, 2013.

[26] Hettige B., Karunananda A.S. & Rzevski G., MaSMT: A Multi-agent System Development Framework for English-Sinhala Machine Translation, *Int. J. Comput. Linguist. Nat. Lang. Process. IJCLNLP*, **2(7)**, pp. 411–416, 2013.

[27] Hettige B., Karunananda A.S. & Rzevski G., Sinhala Ontology Generator for English to Sinhala Machine Translation, *Proc. of KDU International Research Conference*, Colombo, 2014.

[28] Hettige B., Karunananda A.S. & Rzevski G., Multi-agent System Technology for Morphological Analysis, *Proc. of 9th Annual session on Sri Lanka Association for Artificial Intelligence (SLAAI),* pp. 1–7, 2012.

[29] Rzevski G. & Skobelev P. Managing Complexity, WIT Press, 2014.