

Rule Based and Expectation Maximization algorithm for Arabic-English Hybrid Machine Translation

Arwa Alqudsi*, Nazlia Omar*, Rabha W. Ibrahim**

* Knowledge Technology Research Group (KT), Center for AI Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia

** Institute of Mathematical Sciences, University of Malaya

Article Info

Article history:

Received Mar 5, 2016

Revised May 9, 2016

Accepted May 26, 2016

Keyword:

Arabic-English Machine Translation
Expectation–Maximization (EM) Algorithm
Hybrid Machine Translation
Machine Translation

ABSTRACT

It is practically impossible for pure machine translation approach to process all of translation problems; however, Rule Based Machine Translation and Statistical Machine translation (RBMT and SMT) use different architectures for performing translation task. Lexical analyser and syntactic analyser are solved by Rule Based and some amount of ambiguity is left to be solved by Expectation–Maximization (EM) algorithm, which is an iterative statistic algorithm for finding maximum likelihood. In this paper we have proposed an integrated Hybrid Machine Translation (HMT) system. The goal is to combine the best properties of each approach. Initially, Arabic text is keyed into RBMT; then the output will be edited by EM algorithm to generate the final translation of English text. As we have seen in previous works, the performance and enhancement of EM algorithm, the key of EM algorithm performance is the ability to accurately transform a frequency from one language to another. Results showing that, as proved by BLEU system, the proposed method can substantially outperform standard Rule Based approach and EM algorithm in terms of frequency and accuracy. The results of this study have been showed that the score of HMT system is higher than SMT system in all cases. When combining two approaches, HMT outperformed SMT in Bleu score.

*Copyright © 2016 Institute of Advanced Engineering and Science.
All rights reserved.*

Corresponding Author:

Arwa Alqudsi,
Knowledge Technology Research Group (KT), Center for AI Technology (CAIT),
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia.
Email: arwa.alqudsi81@gmail.com

1. INTRODUCTION

There are many languages in the world and it is difficult to find a suitable machine translator that meets human requirements to translate these languages. In the same time translation process faces a number of problems, including linguistic problems, especially in terms of Arabic to English translation, such as word construction, differences in grammar and ambiguous words. Many different approaches have been applied to solve Arabic machine translation problems and its evaluation [4] [5]. These approaches are discussed in detail by [1].

The Rule Based machine translation system is one of these approaches, denotes linguistic rule between source language and target language. Rule based has control on explicit linguistic knowledge that it can totally analyze in both semantic and syntax levels. To create the linguistic rule the Rule Based requires much linguistic knowledge so that it needs high cost development. The result of Rule Based system is depends on the accuracy of each level and the output is less fluency than statistical system. They are many researchers used rule based approach as their translation method, such as: MT for Romance Languages to

Spain [19], Arabic to English [2] and [3], Indonesian to Malaysian [21], Bulgarian to Macedonian [22], English to Sanskrit [23], etc.

In contrast, the Statistical MT is another very common translation system, where probability rules are used, such as, *Baye's* rule of the Expectation Maximization Algorithm (EM). Statistical Machine Translation system is constructed based on parallel corpora. It performs training process on the parallel corpora to learn implied knowledge existing in co-occurrence statistic. Translation for certain word of source language in this system will be found by looking the word in target language which frequently occurs together with them in parallel corpora. Statistical Machine Translation system is able to produce good translations if the source sentence is not similar to any sentences in a training corpus [6].

However, it will have problems in dealing with structures and vocabulary that did not occur in the training data. By using a combined phrase tables from the RBMT systems as well as the SMT parallel corpus table, the hybrid system can handle a wider range of exploit knowledge and syntactic constructions that the RBMT system has about the specific vocabulary of the source text [25]. Linguistic data from RBMT systems have already been used to enrich SMT systems [26-29].

Many researches have attempted to improve this Statistical approach, some used word sense disambiguation [8], and some used word categorization and grammatical categories to handle the error [7]. [9] Used two decoding algorithms to search for the most probable translation of an input tree. Expectation Maximization (EM) algorithm was used to obtain the probabilities from a Treebank. [10] Improved the number of relevant documents retrieved by using EM algorithm and studied the best EM distance for the words in Arabic language that describes the similarity between these words.

If we could combine the advantages of these approaches, the resulting hybrid system could perform better translation than any other system [11]. There are many different types of Hybrid (HMT), since the core of this version of Machine Translation focuses upon a mix of each other. Still a number of researches have been done within HMT. The goal of Hybrid approach is effectively to obtain more accurate results than other existing approaches. The motivation is employing different machine translation paradigms, which implies that a smart combination of their output would return an overall good translation [16]. Translation process of HMT is completed by coupling two or more systems that are employed to solve problems with certain parts. According to present requirements, the most popular combinations comprise RBMT vs. SMT [29].

[12] normalized the dialectal words in a hybrid machine translation (rule based and statistical) system, by performing a combination of morpheme-level mappings and character. They translated the Arabic to English using a hybrid MT. In terms of BLEU system by measuring and comparing the results the author proved the feasibility of the HMT approach. On the other hand the advantages of the rule based and example based approaches used to suggest a form of Japanese-to-English machine translation that can be used with existing technology [13]. Two ways presented by [14] to combine rule-based and statistical approaches to English-German machine translation by integrating existing implementations into a larger architecture.

In this paper we have described the process of developing Arabic-English HMT system as a way to improve the performance of machine translation. We have combined RBMT with SMT using United Nations parallel corpus. The remaining of this paper is structured as follows: Section 2 will describe the architecture of Arabic-English hybrid machine translation system; Section 3 will describe the implementation of HMT; Section 4 will present the experiment result together with its analysis; and Section 5 will give a conclusion about this research.

2. ARABIC-ENGLISH HYBRID MACHINE TRANSLATION SYSTEM

An important trend over the last years lies in a focus shift towards hybrid machine translation systems. The aim of these systems is combining of resources and techniques from different technological backgrounds, e.g., rule based and statistical approaches [24].

Arabic-English HMT system is presented in this paper. This system consists of two main components. Rule based component and Statistical component. RBMT Parser maps the Arabic rules into English rules and SMT techniques handle the language ambiguity using corpus.

2.1. RBMT Component

Rule Based has its origin in transfer system machine translation where it in likeness uses rules [15], and is based on linguistic information about the source and target languages mostly extracted from dictionaries. Rule based machine translation system is a knowing system, because it is based on translation rules rather than a dictionary. When the structure of the source sentence matches one of the rules, it is translated directly using a dictionary. It goes from the source sentence to a morphological analysis and

syntactic analysis to produce a new sort of sentence structure based on rule of the structure source sentence, from this it translates to the target language based on rule of the structure target language and from these steps a better translation is produced to create the final step of the translation.

This research uses our RBMT (AE-TBMT) developed in previous work [2]. Basically, the translation process of AE-TBMT consists of six main phases: first, text in the source language is transferred to tokenizer to divide the text into tokens. Second, start morphological analysis to provide morpho-syntactic information. Third, the syntactic parser builds a syntactic relevant tree, which represents relationships between the words of the phrase. Forth, lexical transfer will map Arabic lexical elements to their English equivalent. It will also map Arabic morphological features to the corresponding set of English features. Fifth, structure transfer will map the Arabic dependency tree to the equivalent English syntactic structure. Finally, Arabic synthesiser will synthesis the inflected English word-form based on the morphological features and traverses the syntactic tree to produce the surface English phrase.

2.1.1. Tokenization

This an important step for a syntactic parser to construct a phrase structure tree from syntactic units. After inserting the source sentence in the system the tokenizer divides the text into tokens. The token can be a word, a part of a word, or a punctuation mark. A tokenizer requests to know the white spaces and punctuation marks.

2.1.2. Morphological analysis

After the tokenization process, the morphological analyser will provide the morphological information about words. It provides the grammatical class of the words (parts of speech) and creates the Arabic word in its right form, depending on the morphological features.

2.1.3. Lexicon

In this system the lexicon is accountable for inferring morphological and classifying verbs, nouns, adverb and adjectives when needed. It is the main lexicon translation; the source language searches in a dictionary and then chooses the translation. A lexicon provides the specific details about every individual lexical entry (i.e. word or phrase) in the vocabulary of the language concerned. Lexicon contains grammatical information which usually have abbreviated form: 'n' for noun, 'v' for verb, 'pron' for pronoun, 'det' for determiner, 'prep' for preposition, 'adj' for adjective, 'adv' for adverb, and 'conj' for conjunction. The lexicon must contain information about all the different words that can be used. If the word is ambiguous, it will be described by multiple entries in the lexicon, one for each different use.

2.1.4. Parsing

The parser divides the sentence into smaller sets depending on their syntactic functions in the sentence. There are four types of phrases i.e. Verb Phrase (VP), Noun Phrase (NP), Adjective/Adverbial Phrase (AP), and Prepositional Phrase (PP). After the parsing process the sentence is represented in a phrase structure tree. Figure.1 show the phrase structure tree for the sentence الرئيس الأمريكي حضر القمة (US President attended the summit).

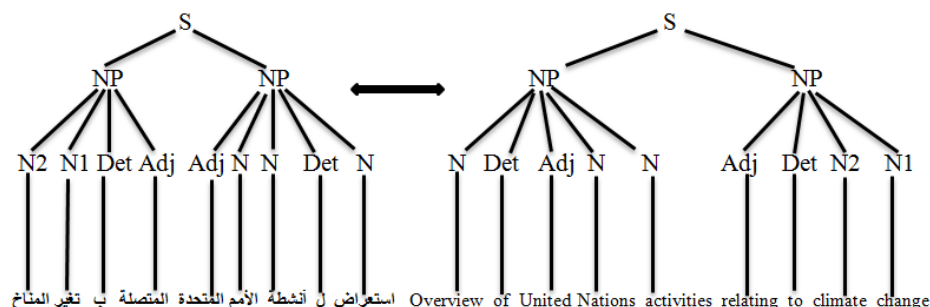


Figure 1. Phrase structure tree

2.1.5. Syntactic rules

A set of Arabic and English rules are fed into the system. In this step the reordering process will be found which will be based on the order of words in a sentence, and how the words are grouped.

2.1.6. Agreement rules

After syntactic rules the agreement rules applied which are responsible about the additions of prefix and suffix in the sentences. Figure. 2 shows an example of Arabic rules and their equivalent in English


S: [NP VP] == S: [NP:[N1 N2] VP:[V P1 N3 P2 N4 Prop]	Arabic rule	
S: [NP VP] == S: [NP: [N2 N1] VP:[V P1 N3 P2 Prop N4]]	English rule	

Figure. 2 Arabic rules and their equivalent in English

There is a significant difference between Arabic rules and English rules. Arabic sentence structure follows Subject-Object-Verb or Subject-Verb-Object or Verb-Subject-Object or Verb-Object-Subject whereas, English follows structure as Subject-Verb-Object. This part is responsible to perform rules matching for the translation. Rule based will match those gaps and translate the sentence. To perform the translation process it needs lots of rules. Adding more rules to the system can enhance the accuracy of the translated output.

2.2. SMT Component

Expectation Maximization Algorithm (EM) is an iterative statistical algorithm for finding maximum likelihood of missed data or missed translations; expect it is the most accurate one, used as the SMT component system in this work.

The Expectation Maximization algorithm contains two processes in each of the iteration: The E-step (Expectation), and the M-step (Maximization). In E-step use current observations and parameters to compute the probability of all possible collocation of the data. This step is achieved by using the conditional expectation.

$$P(C, A | E) = \prod_{j=1}^J t(a_j | e_{c_j})$$

$$P(C | E, A) = \frac{P(C, A | E)}{\sum_A P(C, A | E)} = \frac{\prod_{j=1}^J t(a_j | e_{c_j})}{\sum_C \prod_{j=1}^J t(a_j | e_{c_j})}$$

- A refers to input words (Arabic words)
- E refers to output words (English words)
- J, the number of words in E: $E = e_1, e_2, \dots, e_J$
- c 1-to-many collocation C: $C = c_1, c_2, \dots, c_J$
- For each position in A, generate a word c_j from the collocation word in E: e_{c_j}

In M-step collocation probability which achieved in E-step are used to re-estimate the values of all parameters until converge. The estimate of the missing data from the expectation step is used instead of the actual missing data. Iterating these steps will lead to a convergence that estimates the maximum likelihood of translation. The EM algorithm is discussed in detail by [17]. A common technique for maximum likelihood estimation of model parameters in the presence of missing data is EM algorithm [18].

3. IMPLEMENTATION

The experiments have been conducted on Arabic to English. Hybrid approach relies on structures output by the component Rule Based and Statistical systems.

Rule based machine translation approach generates candidate sentences for each input sentence. Since one word of Arabic can have many meanings in English, thus it is important to identify the suitable meaning of a word for the source sentence. Therefore this generates candidate sentences with each meaning of a word. It creates translated English candidates sentences which are grammatically correct but due to

ambiguous meanings of words there may be meaningless sentences. EM algorithm will choose most suitable sentence from those candidate sentences.

This is the place where SMT techniques come to the work. From Rule based part generate translated candidate sentences for the source sentences based on the number of ambiguous words and the number of ambiguous meanings of each word. Based on the presence of ambiguous meanings of the words, there may be many candidate sentences for source sentence. Then we use a corpus to match most suitable sentence from the given candidate sentences we have used United Nations (Arabic-English) parallel corpus [20] same training and test data split was used as in [31]: 1,000,000 training sentence pairs and tested on 994 test sentences.

Then we use EM algorithm to match other related words to each word in the candidate sentence, by using those related words we create new sentences for each candidate sentence. Then find the probability of each sentence by matching with the corpus.

All of these probability calculations will be considered to select the most suitable sentence. Then we select the correct sentence as the candidate sentence which has highest probability because it has a high possibility to be a meaningful and correct sentence. The architecture of this HMT system is illustrated on Figure. 3.

The data driven RBMT and SMT methods are robust. The feature makes such systems very attractive as they always produce translation, irrespective of the input string. ; If a RBMT system does not find a sequence of rules which can be applied successfully to the input, then the SMT will be identified and produced.

However, statistical system is not good at modelling linguistic phenomena such as word order and agreement. In contrast, RBMT system, can handle linguistic phenomena, such as, word order and using hand-written rules and dictionaries. Thus, the advantage of combining the positive elements of the rule based approach and statistical approach to MT are clear: a combined model has the possibility to be robust, highly accurate, cost effective to build and adaptable. Combining rules with linguistic information and a statistical translation model might result as a hybrid model. The motivations for adopting hybrid model are precisely as mentioned before: it combines the robustness of RBMT and SMT approaches.

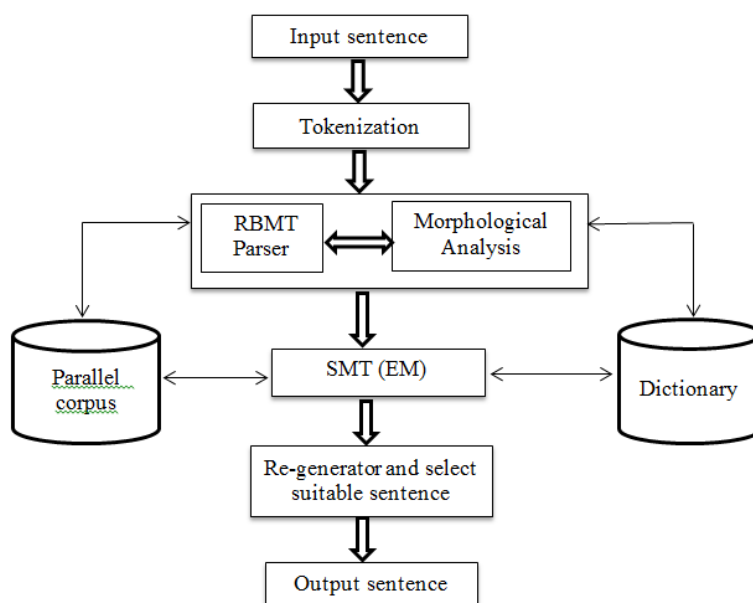


Figure 3. The Architecture of HMT system

4. EXPERIMENTAL RESULT

We have evaluated our system using BLEU [30] scores against two reference human translation. Bleu score for both of the system is calculated and it is described in Table 1. The table shows the values of BLEU obtained for phrase length: *1-gram*, *2-gram*, *3-gram*, and *4-gram*, respectively. Note that BLEU is in

between 0 and 1 ($0 \leq \text{Bleu} < 1$). When BLEU value is close to 1, which means the quality of translation is better and close to the manual translation. In this evaluation 1 candidate file (represent our system translation) and two references files (represent 2 different manual translation) have been used. It can be clearly seen that score of HMT system is higher than SMT and RBMT systems in all cases. When compared three approaches, HMT outperformed SMT and RBMT in all BLEU score.

After performing analysis toward the output of RBMT, we found that, comprehensive reordering rules play an important role in the quality of translation. As HMT system yielded a good improvement in BLEU points over translating of SMT system. In addition, more data training makes the output of SMT more accurate. Figure. 4 illustrates how HMT system translation is closer than SMT system translation to manual translation with phrase length: 1-gram, 2-gram, 3-gram, and 4-gram, respectively. We believe that, a good translation could be achieved when RBMT is combined with SMT, as RBMT solves word ordering problem when translated from Arabic to English, and SMT solves the ambiguity problem.

Table 1 Blue evaluation results of HMT and SMT

Phrase length n-gram	HMT	SMT	RBMT
1-gram	0.88	0.73	0.71
2-gram	0.80	0.61	0.57
3-gram	0.66	0.56	0.50
4-gram	0.51	0.37	0.33

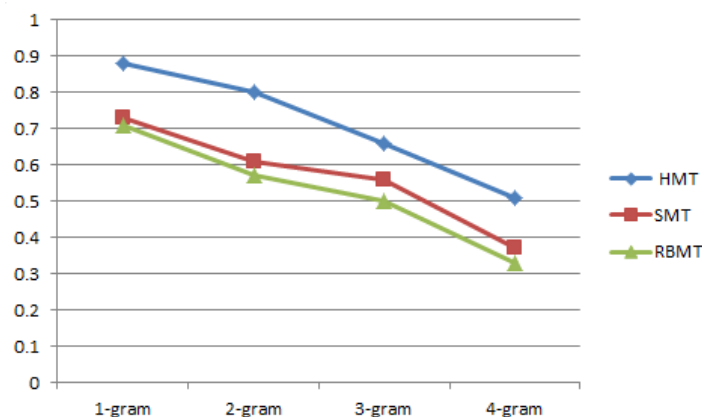


Figure 4. Bleu score of HMT, SMT and RBMT with phrase length 1-gram, 2-gram, 3-gram, and 4-gram.

5. CONCLUSION

In this work we had presented a unique complementary way to combine rule-based and statistical approaches to HMT, as it interweaves the philosophies of the rule based, and statistical approaches in an integrated framework. The goal of combined those approaches effectively to obtain more accurate results than other existing approaches.

This model has the capacity to combine the linguistic complexity of rule based models of translation with the robustness and adaptability of statistical method. The model also helps to address language ambiguity problem which is a one of the biggest challenge under RBMT and solve lexical analysis and syntactic analysis requirement problem in SMT. Thus, we had presented the implementation details of our hybrid system, which was inspired by rule based with the EM algorithm for statistical method; the system has documented larger scale, more translations and complex experiments. This empirical evaluation showed for the Arabic to English United Nations parallel corpus.

The motivation behind this research is combining the advantage of information present in each of the MT system to get better translation result. Evaluation by using Bleu score indicator shows that: 1). The size of the training data effects the statistical model on SMT and HMT system, so adding more training corpus can improve the performance HMT system. 2). HMT system outperforms SMT and RBMT systems in all cases. We had identified that hybrid solutions tend to combine the advantages of the individual approaches to achieve an overall better translation. The approach is most useful to address one of Rule-Based

MT greatest challenges – translation ambiguity. When a word/phrase can have more than one meaning, statistics can help identify the most suitable option.

REFERENCES

- [1] Alqudsi A, Omar N, Shaker K. Arabic Machine Translation: a Survey, Artificial Intelligence Review. 2012: 1-20.
- [2] Hatem A, Omar N. Syntactic reordering for Arabic-English phrase-based machine translation. Database Theory and Application, Bio-Science and Bio-Technology. *Springer Lecture Notes in Computer Science*. 118, 2010: 198-206.
- [3] Hatem, A Omar, N Shaker, K. *Morphological analysis for rule based machine translation, Semantic Technology and Information Retrieval (STAIR)*. International Conference on, 2011: 260-263.
- [4] Arwa Alqudsi, Nazliah Omar, Rabha W, Ibrahim, A New Machine Translation Evaluation Metric Based On The Geometric Mean, (2015). To appear.
- [5] Rabha W. Ibrahim, Arwa Alqudsi and Nazliah Omar, A New Machine Translation Evaluation Metric Utilizing the Holder Mean, (2015). To appear.
- [6] Charoenpornasawat, P Somlertlamvanich, V Charoenporn, T. *Improving Translation Quality of Rule-based Machine Translation*. In: Proceedings of COLING Workshop on Machine Translation in Asia, 2002: 351-356.
- [7] Farrús, M Mariño, JB Poch, M Hernández, A Henríquez, C Fonollosa, JAR Costa-Jussà, MR. Overcoming Statistical Machine Translation Limitations: Error Analysis and Proposed Solutions for the Catalan---Spanish Language Pair. In: *Journal Language Resources and Evaluation*. 2011; 45(2):181-208.
- [8] Carpuat, M Wu, D. *Improving Statistical Machine Translation using Word Sense Disambiguation*. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2005: 61-72.
- [9] Bojar O, M Cmejrek, Mathematical Model of Tree Transformations. ˇ Project Euromatrix – Deliverable 3.2, Prague, Czech Republic 2007.
- [10] Shaalan, K Al-Sheikh, S Oroumchian, F. 2012. *Query Expansion Based-on Similarity of Terms for Improving Arabic Information Retrieval*. Intelligent Information Processing VI. 2012: 167–176.
- [11] Chen, K, Chen, H. *A hybrid approach to machine translation system design*. In Comp. Linguist. and Chinese Language Processing 23, 1997: 241–265.
- [12] Hassan Sawaf. 2010. *Arabic dialect handling in hybrid machine translation*. In Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado.
- [13] Shirai, S, et al. A Hybrid Rule and Example-based Method for Machine Translation, NTT Communication Science Laboratories, pp. 1-5.
- [14] Eisele, A Federmann, C Uszkoreit, H Saint-Amand, H Kay, M Jellinghaus, M Hunsicker, S Herrmann, T Yu Chen. *Hybrid machine translation architectures within and beyond the euromatrix project*. In Proceedings of European Association of Machine Translation (Hamburg), 2008: 27–34.
- [15] S´anchez-Mart´inez, F, Ney, H. *Using alignment templates to infer shallow-transfer machine translation rules*. In *Lecture Notes in Computer Science 4139*, Proceedings of FinTAL, 5th International Conference on Natural Language Processing, 2006: 756–767.
- [16] Federmann, C. *Hybrid Machine Translation Using Joint, Binarised Feature Vectors*. In: Proceedings of the 20th Conference of the Association for Machine Translation in the Americas (AMTA 2012). 2012: 113–118.
- [17] Geoffrey McLachlan and Thriyambakam Krishnan. The EM Algorithm and Extensions. John Wiley & Sons, New York, 1996.
- [18] AP Dempster, NM Laird, DB Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. of the RS Society series B, 1977; 39:1–38.
- [19] Corbi-Bellot, AM, Forcada, ML Ortiz-Rojas, S Perez-Ortiz, JA Ramirez-Sanchez, G Sanchez- Martinez, F legria, I Mayor, A Sarasola, K. *An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain*. In: Proceedings of the Tenth Conference of the European.
- [20] Rafalovitch, R Dale. *United nations general assembly resolutions: A six-language parallel corpus*. Proceedings of the MT Summit XIII, 2009: 292–299.
- [21] Larasati, SD Kuboñ, V. *A Study of Indonesian-to- Malaysian MT System*. In: Proceedings of the 4th International MALINDO Workshop, Jakarta (2010).
- [22] Rangelov, T. *Rule-based Machine Translation between Bulgarian and Macedonian*. In: Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, 2011: 53-59.
- [23] Barkade, VM Deval, PR. English to Sankrit Machine Translation Semantic Mapper. In: *International Journal of Engineering Science and Technology*. 2010; 2(10): 5313-5318.
- [24] Christian Federmann, *Hybrid Machine Translation Using Joint, Binarised Feature Vectors*. In Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas, 2012: 113-118.
- [25] Eisele, A Federmann, C Saint-Amand, H Jellinghaus, M Herrmann, T Chen, Y. *Using mooses to integrate multiple rule-based machine translation engines into a hybrid system*. In: Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, ACL. 2008: 179–182.
- [26] Tyers, FM. *Rule-based augmentation of training data in Breton–French statistical machine translation*. In Proceedings of the 13th Annual Conference of the European Association of Machine Translation, 2009: 213–217.
- [27] Schwenk, H Abdul-Rauf, S Barrault, L Senellart, J. *SMT and SPE Machine Translation Systems for WMT’09*. In Proceedings of the Fourth Workshop on Statistical Machine Translation, StatMT ’09, 2009: 130–134, Stroudsburg, PA, USA. Association for Computational Linguistics.

- [28] Sanchez-Cartagena, VM, Anchez-Mart, FS, Perez-Ortiz, JA. *Integrating shallow-⁺ transfer rules into phrase-based statistical machine translation*. In Proceedings of the XIII Machine Translation Summit, 2011: 562–569, Xiamen, China, September.
- [29] Xuan, H, Li, W, Tang, G. An Advanced Review of Hybrid Machine Translation (HMT). *Procedia Engineering*, 2012; 29: 3017-3022.
- [30] Papineni, KA, Roukos, S, Ward, T, Zhu, WJ. 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109–022), IBM Research Division, Thomas J. Watson Research Center.
- [31] Bangalore, P Haffner, S Kanthak. Statistical machine translation through global lexical selection and sentence reconstruction. In 45th Annual Meeting of the Association of Computational Linguistics, 2007: 152–159, Prague, Czech Republic, June.