Corpus Based Machine Translation for Scientific Text

Irsha Tehseen, Ghulam Rasool Tahir, Khadija Shakeel, Mubbashir Ali

Corpus Based Machine Translation for Scientific Text

Irsha Tehseen[1] (✉) [0000-0001-6969-6323], Ghulam Rasool Tahir[2][0000-0002-4934-6251] , Khadija Shakeel[3][0000-0002-9502-7729] and Mubbashir Ali[4][0000-0003-2467-4045]

[1, 3, 4] University of Lahore, Gujrat, Pakistan
[2] National Language Promotion Department, Islamabad, Pakistan
[1] IrshaTehseen@gmail.com

Abstract. From many years, machine translation and computational linguistic research community has given immense attention towards the development of machine translation techniques. In order to fulfill the goal of machine translation "translation without losing meaning", a lot of translation methods have been proposed. All of these translation methods differ in their theories and implementation strategies. Although some basic rules of translation are same but many of them vary with the selection of language pair. While concerning with the scientific text, every science domain has thousands of terminologies. Translation of these terminologies according to the domain boosts the performance of translation. Translation of scientific text is ignored in the literature, as it needs more effort and expertise of both domain and language are required. In this research, we have proposed an effective scientific text translator for English to Urdu to cope with the challenge of scientific text translation. This method tags and translate the terms according to the domain. We have introduced a term tagger for tagging terms. The system can work for any domain but for experimental purpose we have selected the domain of computer science. System is evaluated on self-generated corpus of computer science. It is also compared with the existing translators to demonstrate the dominance of proposed translator as compared to the competitor. The comparative results of proposed approach and existing are shown in the form of tables.

Keywords: Machine Translation, Corpus Based Machine Translation, Scientific Text Translation, Term tagger

1 Introduction

Language is the main medium for humans to communicate. Whenever humans need to communicate, having different languages, they have to face issues. This arises the demand to translate. This demand is as old as the human [1]. Human experts in multiple languages are offering their services of translation from many decades. The demand of translation is increasing with the growth of cross-regional communication. Different sources of data are accessed worldwide. They cannot be written or translated in every language manually. They need to be written in one language and automatically translated in user preferred. This demands to cover the barrier of the language. [2]. Due to increase of translation demand humans are not able to fulfill the needs of the society,

in response the automate process of translation is generated [3]. This idea fascinated the researchers. This research area is known as Machine Translation.

1.1 Machine Translation

Machine Translation (MT) is the process of translation from one language to another by the use of computing devices [1]. MT is an automatic process in which all the translation jobs are done with the help of programming languages and software [1, 4]. All materials needs to be translated, this includes commercial, business and scientific documents, instruction manuals, text books, World Wide Webs. [3, 5].

Machine Translators are serving for multiple languages. Quality of MT also varies with language pairs. Two languages for instance English to French may have high quality of translation, some other pair may have translation of low quality. Currently none of the language pairs are translated accurately. The main difficulty in automated translation of one language to another is varied written scripts and multiple lexical choices for a single idea. A single world may have various meanings, in different situations it is used for different purposes. At the current level of research, a single level of representation of every language is almost impossible. Every language needs to be considered separately with its scripting choices for automatic translation [6].

1.2 English to Urdu Machine Translation

English and Urdu both are Indo-European languages but differ in written scripts and morphology. Urdu is a right to left scripted language and Eng. is left to right. Eng. follows the same order while Urdu is a free order language. English always follows the subject-verb-object order, Urdu mostly follows subject-object-verb pattern but not always [7]. Although a lot of work is done for MT but still English to Urdu MT is in its early stage. This pair of language is considered a low resource language because enough standard translated text is not available for the training of the system [8].

1.3 Machine Translation Service for Scientific Text

The number of scientific texts other than English keeps increasing quickly as compared to past, as the scientific communities in non-English countries grow [9]. However, majority of high impact journals are published in English [9]. For translating scientific text, considering only the semantic representation is not enough.

Different terminologies have various meanings in multiple domains. A single word may give a total different concept in various fields. Terms are different in every subject, while translating from one language to another these terms should be considered according to the scenario. For the true sense of data all of the terms should be translated with true meaning of the domain. None of the translator is working on the translation of scientific text in true meaning. While translating scientific text accuracy is a major issue [10]. We can get more advantage by using MT for translating these scientific text in local language or from local language to English language.

1.4 Problems in Automated Translation of Scientific Text

Translation of scientific text is not as simple as it seems. There are approx. 10 main branches of science and each branch is followed by many sub branches so approx. there are more than 100 fields of sciences [11]. Each field has different terms and meaning of terms, so a single generic system for all these sciences is not easy.

Translation of scientific text requires domain and translator expert. These texts are written by using Languages for Special Purposes (LSP). Translation of scientific text not only requires the knowhow of the language but it also requires the deep understanding of the field. For the translation of scientific text both of the skills: translation skill and domain skill are compulsory [12, 13].

Although there are many translators doing the job but translation of scientific text is still ambiguous. Translators are not trained for domain skills. Translators are working only with the translation. There is need to develop a MT system which can be a benchmark for translating scientific text while considering domain of the text.

2 Shortfalls in Existing MT Techniques for Scientific Text

Various terminologies are overlapped in multiple fields but their meaning is different in ever field. Such as the word ”monitor” is a device in the field of Computer Science (CS) , but in classroom environment it is used for a student. The word ”Python”, in CS is the name of a programming language, outside of the CS it is considered a snake. Both examples shows generic words may have different meanings in various field.

Existing translators are translating data in generic meanings. They are not able to translate scientific text in real sense. These systems do not distinguish between domains. Above mentioned are the few samples of basics sentences of CS. Translating a whole book or a research paper is more pathetic, humorous.

So far no such benchmark is available for translating scientific text. People have to make extra effort in understanding scientific text because have to understand the language also. So there is need of a study, to identify different techniques which should be used for better translation of scientific text. It is to decrease the effort in learning language skills or translating manually. Existing work do not bridge this gap to translate the scientific text with correct sense. There is a need to develop a customized translator which can effectively bridge this gap. Domain specific translators are tend to give better results as compared to the translation of generic systems [14, 15, 16].

3 Methodology

Terms of any field plays an important role in the translation of the text. It enables to understand the meaning or idea correctly. For quality translation a term tagger can play an important role in translation. It is complex to create single term tagger for all the fields together. This section focuses on the development of a scientific text translator and a term tagger for the field of CS. This translator and term tagger can be trained for any domain of science. The main contribution of the proposed work is term corpus of

CS, translation of that corpora and, a term tagger of scientific text and translation of text according to the meaning of the domain.

3.1 Overview of Proposed Scientific Text Translator

The development of a quality translator is a challenging and tricky task in MT research community. This is mainly due to the diversity of the languages. We have proposed a customized domain specific translator for scientific text. A complete overview of the system is given in the Fig.1. The process of generating scientific text translation is comprised of following steps:

Proposed Algorithm Overview:

1. Check number of sentences entered
2. If sentence is more than one, separate each sentence
3. Create a list of sentences
4. Select a sentence
   a. Remove special characters, symbols, white-spaces and tabs
5. Check for term in the sentences
   a. If term found, tag the term and also generate phrases based on tagged term
      (1) Repeat step 5a for all the terms in the sentence
   b. If linked word found, generate phrases based on linked words
      (1) Repeat step 5b for all the linked words in the phrase
6. List of phrases generated
7. Pick a phrase, if phrase is tagged as a term; search phrase in term base
   a. Retrieve term case b. Repeat step 7 for all the terms
8. If phrase is not a term, check in case base
   a. Compute similarity of the phrase in case base
   b. If similarity is 1, retrieve the case c. If similarity is less than 1
9. search for most similar case
10. Retrieve most similar case
11. Repeat step 8 for all the non-term phrase cases
12. Generate list of phrases
13. Reorder the retrieved cases as
14. Repeat step 4 to 11 for all the sentences and present solutions of the reordered case

It is composed of 4 modules. In module 1, inputted text is converted into plain text. The module 2 tag the terms and divide the sentences into phrases. CBR (Case Based Reasoning) Trainer is used for the searching and retrieving case from case base in 3rd module. At the end, module 4 is used to reorder the phrases to make the translation quality a bit better and readable.

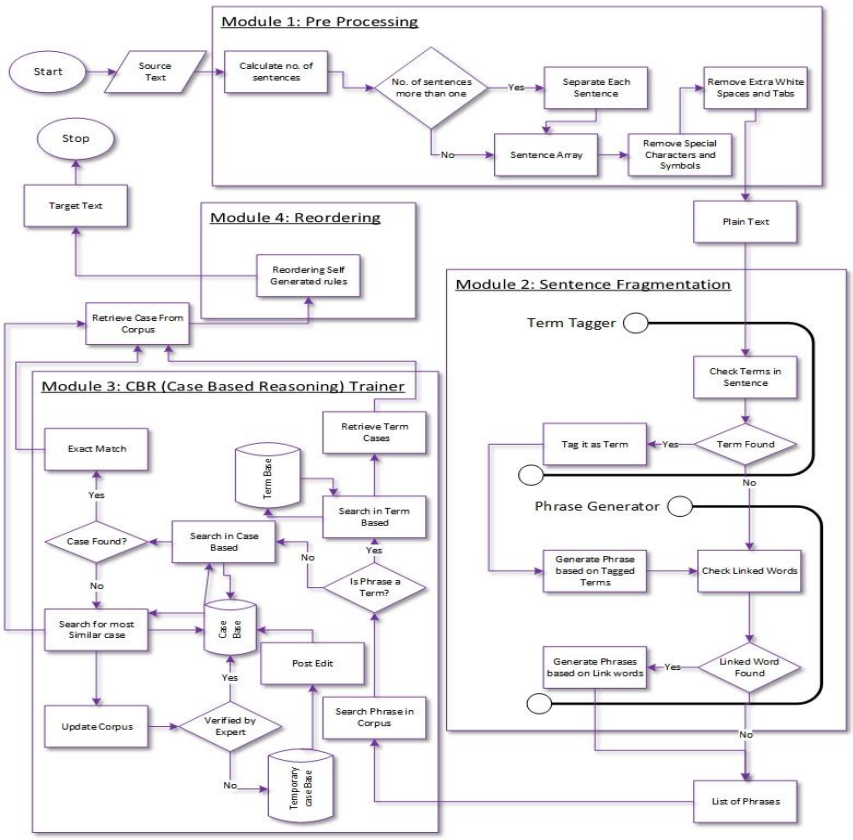

Fig. 1. Overview of Proposed Scientific Text Translator

3.2 Module 1: Preprocessing

All the formatted characters, special numbers, tags, images are removed from the text. At the end of this step text is totally normalized, only readable English characters is present in the text. Translation of any character, special symbol or syntax is ignored.

Preprocessing Algorithm: Input: English Text Output: Plain text of English

Check no. of sentences entered

If sentence is more than one separate each sentence

For each sentence, in the list of sentences

If special character or symbol is found

Remove special characters, symbols, and whitespaces If ends Loop ends

3.3 Module 2: Sentence Fragmentation

There are two ways to keep the sentences into the corpus. One is to keep whole sentence. It will decrease the scope of the sentence as one sentence is equal to only one

example. The second choice is to fragment it into multiple phrases. Each sentence is divided into two or more phrases. Scope of the sentences is increased with this. A broader range of sentences is covered by using genetic algorithm [17]. This module is divided into the following submodules: Term Taggers, Phrase Generator.

Term Tagger: Terms are used to express a concept, mainly in a particular domain of the study. A list of terms is developed to handle the terms of the computer science. Term Tagger tag and checks whether the terms are present in the sentence or not. If it found one or more terms in the sentence, it tags the terms are T1, T2.....Tn

Phrase Generator: Generating phrases is a required and vital module [18]. More phrases leads to more accurate results. Our phrase generator is based on tagged terms and linking words. Linking words are available at [19]. These words are used for further fragmentation.

Fragmentation Algorithm: Input: English sentence Output: Set of English phrases
For each term, in the list of terms
Find term in the input sentence
If term is found tag it as a term
Separate the sentences into new sentences If ends Loop ends
Find linked words
For each linked word separate the fragments of sentences

3.4 Module 3: CBR (Case Based Reasoning) Trainer:

CBR Trainer is responsible of searching, measuring similarity and retrieving the solution of new case based on the old cases or training.

Searching in Corpus: Searching is checking whether the input phrase is available. If exact match is available its translation is presented. If it's not available the most similar solution is presented. Similarity of the case is checked in two ways: exact match and most similar case.

Searching Algorithm: Input: Phrase Output: Case ID
For each phrase in the list of phrases check it is tagged as a term or not
If phrase is a term
Compare the term in the term base and return ID of the term case
Else
Compare the phrase in case base
If exact match found return matched case ID from case base
Else
Search most similar case in case base
If similarity of the case is ≥ 0.8 return the ID of the selected case
Else
Tag the case as approx. solution
add the case in update corpus and return the ID of selected case
If ends If ends If ends Loop ends

Retrieving from Corpus: The exact match or the most similar case is retrieved and its corresponding translation is presented. If a sentence is based on a single phrase its translation is presented directly. If sentence is based on two or more than two phrases,

translation of every phrase is retrieved individually. Later on they are combined to formulate a single sentence.

Retrieving Algorithm: Input: Case ID Output: Solution of all cases in Urdu
Set a=1 For each case en_a
If case is a term case
Find the case en_a in the term base and make a set Ur_a
Find the translation solutions of en_a and add it to Ur_a and add 1 to a
Else
Find the case en_a in the case base and make a set Ur_a
Find the translation solutions of en_a and add it to Ur_a and add 1 to a
If ends Loop ends Find the union of all Ur_a's

Note: *a* is a counter variable for counting number of phrases
en is a set of English Computer Science (CS) phrases
ur is the set of Urdu CS Translation phrases
for each e_a there is an equivalent part Ur_a. It is considered as follows:
En = Set of English CS phrases = { en_1, en_2, en_3.....en_n }
Ur = Union of all the Urdu CS Translation phrases = $Ur_1 + Ur_2 + Ur_3$.....Ur_n

Here number of phrases of a particular sentence is not specific and cannot be known before the actual program executes. These phrases are constructed at the run time. Solution of the cases are also only available at the run time.

Updating Corpus: New solved cases are saved for future use. Case base is updated but these cases are kept separate until they are post-edited and verified by expert.

3.5 Module 4: Reordering:

Union of *Ur* is presented as output. Reordering of the sentence is a separate issue. Here we only consider it to some extent, just to make the translation a bit readable.

Reordering Rules: If en_1.....en_n are the CS phrases in English whose equivalent phrases are $Ur_1 + Ur_2$....Ur_n, then the translation of [en_1, en_2, en_3.....en_n] is [$Ur_1 + Ur_n + Ur_{n-1}$..... + $Ur_3 + Ur_2$]

4 Experimental studies

Here we present self-generated corpora, its translation and results. The accuracy results of our system are presented and compared with existing translators.

4.1 Experimental Corpus:

We used self-generated corpus. Generating corpus is a weighty research extraction. There are two corpus: Term Corpus; Base Corpus

Term Corpus: It is our first corpus, for this corpus the terms are picked from multiple sources [3, 21, 22, 23, 24]. Many resources have same terms, overlapped terms are cleaned and discarded automatically, later they are checked manually.

Base Corpus: It is our second corpora. It consists on CS phrases. These phrases are constructed from multiple sentences. Sentences are selected from CS books, research papers and Wikipedia page of CS. These sentences have various length and terms in it. Sentences are fragmented into phrases by using proposed fragmentation algorithm. All the duplicate phrases and special symbols are removed.

Translation of Corpora: It is another major issue. As there are very few standard translated text of CS is available in Urdu language. Text is translated as accurate as it can. Translation of term requires a careful and persistent effort [25]. How we translated these terminology is also a separate issue. Translation of Term corpora is done according to the meanings of CS and it is also revised twice. The second step is translating Base corpora into equivalent Urdu translation. These translations still can be improved by expert. A concise overview of the above explained corpora is given in the Table 1.

TABLE 1. A concise overview of our experimental datasets

Sr#	*Corpora*	*Dataset Description*	*Total*	*Unique*
1	Term Base	A Bilingual Corpus (English to Urdu) of unique CS terms	14002 Terms	10156 Terms
2	Case Base	A Bilingual Corpus (English to Urdu) of unique CS text phrases	1500 sentences / 18982 Phrases	14232 Phrases

4.2 Experiments:

In this section, we presented the experiments to evaluate the performance of our system. These experiments are performed by using datasets discussed in Table 1. The accuracy results of our proposed system are shown in Table 2.

Experiment 1: Evolution of Proposed Scientific Text Translator:

The purpose of this experiment is to evaluate the translation accuracy of proposed system. The experiment has been conducted on above mentioned datasets. After giving the text, first step performed is preprocessed the text according to the algorithm proposed. Second step performed is to tag the terms in the given sentence and generate phrases fragmentation algorithm. Third step is to search and retrieve solution cases. At the end of the translation process reordering is done.

Experiment 2: Comparison of Proposed Translator with competitor:

Purpose of Experiment:

The aim to conduct this experiment is to compare the proposed system's accuracy with existing systems. The experiment is performed on our internally generated datasets. This experiment is performed in three different steps. In first step, 500 sentences of CS are selected from the corpora and CS books, 50% sentences are selected from corpus, 25% from different text books and rest 25% is from Wikipedia page of CS. Second step is to check translation of selected sentences one by one and verify those translations on famous existing systems and analyzed how much of them are translated correctly.

Comparison of Existing and Proposed Scientific text translation:
Experimental results of competitors for scientific text translation are given in Table 2. We can observe that the existing systems are giving very less translation accuracy and very less terms are translated correctly. These sentences are tested on different famous translation systems. Results are given below in the table. It can be clearly seen that proposed system gives more accuracy as compared to the existing system.

Table 2. A concise overview of our experimental datasets

#	Translator	Type	Sentences test	Terms in sentences	Successfully translated terms	Accuracy %
1	Proposed System	Domain Specific	500	379	301	79%
2	Google	Generic	500	379	102	26%
3	Being	Generic	500	379	153	40%
4	Babylon	Generic	500	379	77	20%
5	Dictionary Translator	Generic	500	379	46	12%

5 Conclusion

We have introduced an effective scientific text translator. The proposed translation method is based on tagging terms of sciences and corpus based MT approach by using CBR. To meet the challenges of terminologies translation of scientific text, a term tagger for scientific text is proposed. It tags the terms of sciences and then translate with the help of self-generated Term Corpora. The performance of proposed technique has been evaluated by performing experiment on self-generated English to Urdu parallel bilingual dataset of CS. Both corpus are developed and translated. Experiment has also been conducted to provide a comparison between proposed technique and existing translation services. From the comparative results we concluded that, the proposed translator accuracy results are significantly better as compared to existing translator approaches. It gives considerable accuracy rate. Our proposed technique is also capable of handling other fields of sciences, all we need is to train the system for that domain. The current training of the system is done on the domain of CS. If we change its training, it can effectively work for every domain of life.

References

1. A. H. Homiedan, “Machine translation,” Journal of King Saud University, 1998.
2. C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan, “Findings of the 2011 workshop on statistical machine translation,” in Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 22–64, Association for Computational Linguistics, 2011.
3. W. J. Hutchins and H. L. Somers, An introduction to machine translation, vol. 362. Academic Press London, 1992.

4. S. Khan and R. Mishra, “Translation rules and ann based model for english to urdu machine translation,” INFOCOMP Journal of Computer Science, vol. 10, no. 3, pp. 36–47, 2011.
5. N. A. Khan, L. Ansari, S. R. Mahmud, M. Sultana, A. Muntaheen, and M. N. Huda, “Bangla to english machine translation,”
6. I. Garcia, “Beyond translation memory: Computers and the professional translator,” The Journal of Specialised Translation, vol. 12, no. 12, pp. 199– 214, 2009.
7. B. Jawaid, A. Kamran, and O. Bojar, “English to urdu statistical machine translation: Establishing a baseline,” COLING 2014, p. 37, 2014.
8. K. M. A. Salam, S. Yamada, and T. Nishino, “Example-based machine translation for low-resource language using chunk-string templates,” 13th Machine Translation Summit, Xiamen, China, 2011.
9. P. G. Altbach, “The imperial tongue: English as the dominating academic language,” Economic and Political Weekly, pp. 3608–3611, 2007.
10. M. Olohan and M. Salama-Carr, Science in Translation. Taylor & Francis, 2014.
11. G. Sandstrom, “How many ‘sciences’ are there?,” Social Epistemology Review and Reply Collective I, vol. 10, pp. 4–15, 2012.
12. S. E. Wright and L. Wright, “Editors’ preface: Technical translation and the american translator,” Scientific and technical translation, Amsterdam/Philadelphia, John Benjamins, pp. 1–7, 1993.
13. J. Byrne, *Scientific and Technical Translation Explained*. Taylor & Francis, 2015.
14. J. C. Micher, “Improving domain-specific machine translation by constraining the language model,” tech. rep., DTIC Document, 2012.
15. J. Xu, Y. Deng, Y. Gao, and H. Ney, “Domain dependent statistical machine translation,” in MT Summit, 2007.
16. B. Hatim and I. Mason, Discourse and the Translator. Routledge, 2014.
17. H. Echizen-ya, K. Araki, Y. Momouchi, and K. Tochinai, “Machine translation method using inductive learning with genetic algorithms,” in Proceedings of the 16th conference on Computational linguistics-Volume 2, pp. 1020–1023, Association for Computational Linguistics, 1996.
18. J. Mallinson, R. Sennrich and M. Lapata, "Paraphrasing Revisited with Neural Machine Translation", in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017.
19. “Linking words and phrases.” https://www.dlsweb.rmit.edu.au/lsu/ content/4_writing-skills/writing_tuts/linking_LL/linking3.html. [Online; Accessed: 26- June- 2016].
20. D. Oțăt, "Corpus-Based Training to Build Translation Competences and Translators’ Self-Reliance", *Romanian Journal of English Studies*, vol. 14, no. 1, 2017.
21. K. Knight, “Machine translation glossary.” http://www.isi.edu/ natural-language/people/dvl.html. [Online; Accessed: 02- May2016].
22. H. Henderson, Encyclopedia of computer science and technology. Infobase Publishing, 2009.
23. P. Koehn, Statistical Machine Translation. Statistical Machine Translation, Cambridge University Press, 2010.
24. M. Press, Microsoft Computer Dictionary. CPG Series, Microsoft Press, 2002.
25. M. Olohan, *Scientific and technical translation*. 2016.