Translation Rules for English to Hindi Machine Translation System: Homoeopathy Domain

Sanjay Dwivedi and Pramod Sukhadeve
Department of Computer Science, Babasaheb Bhimrao Ambedkar University, India

Abstract: *Rule based machine translation system embraces a set of grammar rules which are mandatory for the mapping of syntactic representations of a source language, on the target language. The system necessitates good linguistic knowledge to write rules and require of acquaintance source such as corpus and bilingual dictionary. In this paper, we have described the grammar rules intended for our English to Hindi machine translation system to translate the homoeopathic literatures, medical reports, prescription etc. The rules which have been written follow the transfer based approach for reordering of rules between two languages. The paper first discusses about our developed stemmer and its rules, further we discuss the Part of Speech tagging (PoS) rules for categorizing each word of the sentence grammatically and our developed homoeopathy corpus in English and Hindi of size 20085 and 20072 words respectively and at the last we discuss the agreement/translation rules for translating various homoeopathic sentences*.

Keywords: *Machine translation, stemmer, PoS tagging, grammar rules, homoeopathy, corpus.*

Received June 14, 2013; accepted March 17, 2014; published online September 15, 2015

1. Introduction

Rule-based machine translation systems rely on innumerable built-in linguistic rules and data sources such as bilingual dictionaries for each language pair. In such system, translation rules have various techniques; some of these are entirely automated while others necessitate a lot of human input. The primary step towards the development of translation rules for any language demands an in-depth understanding and analysis of language.

A number of MT systems have been developed for translation of contents using different domains like tourism, story books, newspaper, etc. In the medical related domain, CliniTrans [5] is a famous machine translation system; it provides certified translation and interpretation for the health professions. It translates medical reports and Clinical forms, pharmacy and medical device marketing. In the specialized medical domain such as homoeopathy, however, no significant work has been reported so far. Homoeopathy is idiosyncratic than other clinical languages like allopathy, Ayurveda, etc., it is different in case of medicines, prescription and in case of collecting symptoms from the patients. This paper, focuses on a rule based MT system for this domain and presents translation rules which have been developed according to homoeopathy clinical sentences. These rules are used to translate various documents such as homoeopathic literature, doctor's prescription, medical reports etc., from English to Hindi language.

We perceive certain rules for sentence analysis and translation, through stemming and PoS tagging for both English and Hindi languages. For this purpose, we have discussed our stemmer, a module that finds the root word, by stripping away the affix attached to the word. Stemming is a widespread form of language processing in most information retrieval systems [8, 10]. It is similar to the morphological process used in natural language processing, but has somewhat different aims. It also, helps in clinical language for knobing the clinical terms such as the name of the deceases, patient's medical report and symptoms of the patient. The stemmer for target language (Hindi), which is morphologically very rich, having different morphological variants of a single root word, like औषधियों{aushadiyoon}, औषधियाँ{aushadiyan}] of the root word [औषधि] [17] has also been developed. Stemming rules for homoeopathy clinical language in English and Hindi languages are described in section 4. Next, the Part of Speech (PoS) tagger has been discussed. Tagging is a process of assigning precise syntactic categories to every word in a sentence as corresponding to a particular part of speech, based on its definition, as well as its context [9]. A number of PoS taggers have been developed such as: "The stanford natural language processing tagger" for English [18], "CLAWS part-of-speech tagger" for English [4], "twitter PoS tagging" [19]. PoS tagging for clinical language have gained an increased interest over the past few years, yet the lack of availability of annotated corpora resources obstructs the research and investigations, standardization is another problem because so far no standard tag sets are available for such languages. We have therefore developed our own tag set for the process of tagging and techniques [6]. It has been developed via linguistic rules. Further, the translation rules have been discussed in the section 5.

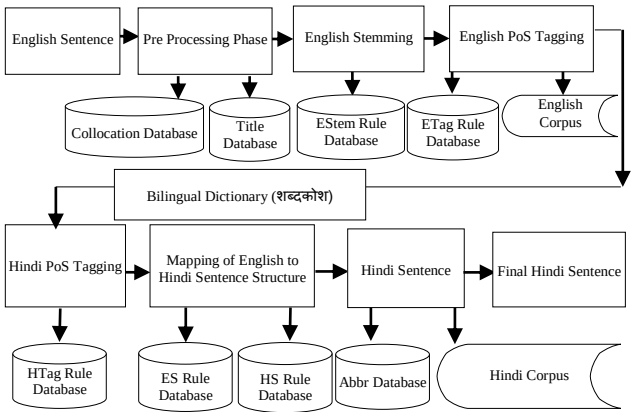
Section 6 discusses translation/ agreement rules for translating homoeopathic sentences from English to Hindi language followed by results and discussion in section7.

2. Related Research Works

Several researchers have worked on machine translation systems and many measures and methods have been developed for well designed systems and a number of literatures are available reporting the development of rule based systems for Indian languages. In the work of Bahadur *et al*. [2] a target language generation mechanism has been outlined with the help of English to Sanskrit language pair using rule based machine translation technique. This system is named as "Etrans". The result of this system as claimed by the authors is excellent as it translated ninety percent of the sentence correctly out of 500 sentences. Another system, MT from English to Bangla [7] language translation model which relies on rule based methodologies especially fuzzy rules was proposed and the authors portrayed "If-Then" basis rules apply for English to Bangla language translation. A knowledge representation technique is used to classify each English sentence to a particular class using attributes of that English sentence and then translate them to the Bangla sentence using the rules and also formed English to Bangla bilingual dictionary for language translation. In the work of Batra and Lehal [3] has been proposed a rule based machine translation for noun phrases for Punjabi language, it has been trained with 2000 phrases and the accuracy has been 75-85%. Angla Malayalam system from CDAC is developed for health and tourism domains with 75-80% accuracy. AnglaBharti Technology for machine aided translation from English to Indian languages [14] is another prominent work.

3. System Architecture

The architecture of the Homoeopathic MT system is shown in Figure 1. It is a rule based system. The first step (Pre processing phase), is a collection of operations that are applied to input text to make it processable by the translation engine, in this step various activities incorporated include replacing collocations and titles, which are stored in the respective databases (collocation database and Title database). In the second step (English stemming), stemming rules are applied to get the root words which are stored in the rule database (EStem rule). Further, system tags all the root words obtained in step 2 with the help of PoS tag rules stored in rule data base (ETag rule) and the English corpus. While tagging, if the system finds more than one tags for a single word, it identifies the correct sense of the word according to sentence structure using English corpus. The online bilingual dictionary (शब्दकोश) [13] is used for obtaining Hindi words corresponding to the English words. The system considered only the fine grained meaning of English words. It is possible that the first meaning which the system opts is not suitable for the Hindi sentence. In this case, it resolves the problem using Hindi homoeopathy corpus. After obtaining the Hindi words, system tags all the Hindi words using tag rules stored in the rule database (Htag rule). Further, system maps English and Hindi rules using the rule databases (ES rule) and (HS rule) respectively. The mapped rules are stored in another rule database (assembly rules). After mapping the rules, system generates output (Hindi sentence), which are further matched with the Hindi corpus to remove any possible ambiguity.



Whereas, EStem rule database-English stemming rule data base , ETag rule database-English tagging rule data base , HTag rule database-Hindi tagging rule data base, ES rule data base -English sentence rule database, HS rule database-Hindi sentence rule database.

Figure 1. Architecture of homoeopathic MT system.

4. Stemming Rules for English Sentences

Stemmers are used to convert inflected words into their root or stem word. Stemming improve performance by reducing morphological variants into same words. As the existing stemmers are not able to stem some of the homoeopathic words, a stemmer [15] has been developed on windows platform with Graphical User Interface (GUI). In homoeopathy, there are lots of rules for suffix and prefix of the words; some of the designed rules are:

- *Rule* 1: Replace "ies" by "y".
- *Rule* 2: Remove prefix "dis".
- *Rule* 3: Replace "ing" by "e".
- *Rule* 4: Remove "s".
- *Rule* 5: Remove "ly".
- *Rule* 6: Replace prefix "hyper".

The process of stemming can be explained using the following example:

- *Example* 1: Calc. Sulp. can be excellent remedies in making the boils liberate and clear quickly.
- Initially: Detach all the words from the given sentence to find out the root word.

In the above sentence, we scan all the words and eliminate affixes from the particular words like "remedies, making, boils, quickly", the word "remdies" is stemmed to "remedy" using rule 1, the next word is "making" is stemmed to "make" using rule 3, the next word "boils" is stemmed to "boil" using rule 4 and at last "quickly" stemmed to "quick" by using rule 5. After scanning all the words we get,

- Output: Calc. sulp. 200 can be excellent remedy in make the boil liberate and clear quick.

5. PoS Tagging Rules

The connotation of POS for language processing is the large amount of grammatical information give about a word. There are principally two approaches to PoS tagging: rule based tagging and stochastic tagging, Al-Taani and Abu Al-Rub [1] proposed tagging system which classifies the words in a non-vocalized Arabic text to their tags through rule based approach. Figure 2 shows the steps for analyzing sentences for tagging.

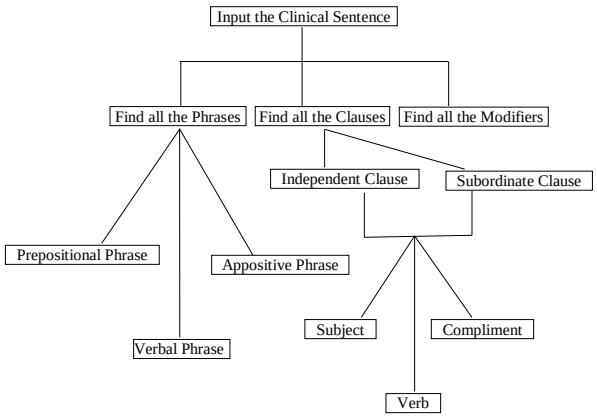


Figure 2. Steps for sentence analysis process.

First we find out prepositional phrase like {aboard, above, behind, despite, past, upon, with, etc.,} and compound preposition like {according to, in addition to, in spite of, instead of, out of, in place of, in regard to, because of, etc.,} from the list of phrases, which are commonly used in the sentence. Next we find out all the clauses like independent clause and subordinate clause like subject, verb and compliment and next we find out all the modifiers like adjectives, adverbs.

- English PoS Tagging Rules: We are using plenty of rules for grammatical tagging of all the words from the sentences and also use our developed homoeopathic corpus of the English language, as of now, 256 rules have been designed. Some of the rules have been discussed below:
  - *Rule* 1: A noun can often be identified by the words around it (noun signals/determiners).

    a. Articles: A, an, the, etc.
    b. Possessives: 's, s', my, our, etc.
    c. Numbers: Five, 23, two, etc.
    d. Indefinites: Some, many, several, etc.
    e. Demonstratives: This, that, these, those, etc.

  - *Rule* 2: A noun very often comes at the beginning and is the subject of a sentence.
  - *Rule* 3: A noun can come after an action verb and is the direct object of the verb.
  - *Rule* 4: A noun can come after the verb and be the indirect object of a verb.
  - *Rule* 5: A noun can come after a preposition and be the object of a preposition.
  - *Rule* 6: Prepositional phrases between the subject and verb usually do not affect agreement. Example: The colors of the stool are different.
  - *Rule* 7: The words that come between the subject and verb; they do not affect agreement. Example: The doctor, who is treating me, is usually very good.
  - *Rule* 8: When sentences start with "there" or "here," the subject will always be placed after the verb.
  - *Rule* 9: Subject don't always come before verbs in questions.
  - *Rule* 10: If two subjects are joined by and, they typically require a plural verb form.
  - *Rule* 11: The verb is singular if the two subjects separated by and refer to the same person or thing.
  - *Rule* 12: if one of the words each, every or no comes before the subject, the verb is singular.
  - *Rule* 13: If one subject is singular and one plural and the words are connected by the words or, nor, neither/nor, either/or and not only/but also.

We employ 125 sentences (2322 words) collected randomly from 20085 words corpus of homoeopathy. Only four diseases have been taken from the complete corpus for tagging accuracy of the tagger, it is computed using the following formula [6]:

$$Accuracy = \frac{Correctly\ tagged\ words}{Total\ number\ of\ tagged\ words} \quad (1)$$

The outcome was manually appraised to mark the correct and incorrect tag assignments. Table 1 shows the total number of words taken for each of the four diseases for tagging and number of correctly and incorrectly tagged words. The same result has been shown in Figure 3.

Table 1. Performance of POS tagger.

Diseases	Incorrectly Tag	Correctly Tag	Total Words
Rheumatism	28	421	449
Anaemia	58	735	793
Migraine	30	130	160
Keloids	110	810	920
Total	226	2096	2322

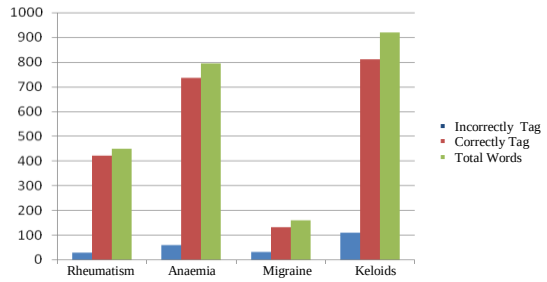


Figure 3. Accuracy graph of PoS tagging (English).

As an example, the clinical sentence "Calc. sulp. 200 can be excellent remedies in making the boils liberate and clear quickly." is tagged as: Calc. sulp._(N) 200_(NUM) can_(VERB) be_(VERB) excellent_(ADJ) remedy _(NOUN) in_(PREP) make_(VERB) the_(VERB) boil_(VERB) liberate_(VERB) and_(CONJ) clear_(ADJ) quick_(ADJ).

- Hindi PoS Tagging Rule: Hindi PoS tagging [16] is another step for translation, for this we use grammar rules and corpus for tagging and also, use our developed homoeopathy corpus in Hindi language. Some of the rules are discussed:
  - *Rule* 1: Most of the feminine nouns in Hindi usually end in आ (aa), इ (i), इया (iyaa), इन (in), आनी (aani), आइन (aain), इका (ika), नी (ni), अती (ati), वती (vati), त्री (tri), etc.
  - *Rule* 2: The verb combined with a feminine noun usually end with इ (i) such as खाती (khati), पीती (piti), सोती (sauti), करती (karti), जाती (jati), etc.
  - *Rule* 3: The masculine verb have a similar pattern, it ends आ (aa) such as खाता (khata), पीता (pita), सोता (sauta), करता (karta), जाता (jata), etc.

  Using the rules and hindi to English bilingual dictionary, we obtain an कैलकेरिया सल्फ्यूरिका (calkeria sulphurica) (NOUN) 200(NUM) सकना (sakana) (VERB) होना (hona) (VERB) अच्छा (achha) (ADJ) इलाज (iilaaj) (NOUN) में (may) (PREP) करना (karna) (VERB); यह (yah) (PREP) वह (vah) (PRON) छला (chala) (NOUN) मुक्त (mukat) (ADVERB) करना (karna) (VERB) साफ (saaf) (NOUN) जल्दी (jaldi) (ADVERB) A.

6. Translation Rules between English and Hindi Language

For language translation from English to Hindi, we need to converse a comparative structure between these two languages. First, we illustrate the sentence patterns of these two languages. Next, we explaine the grammar patterns of the two languages. Later, we compared these patterns for both English and Hindi.

The sentences may be categorised as: Simple sentence, complex sentence and compound sentence. Compound sentence can also, be divided into double and multiple sentences. For example, a simple sentence for English pattern (subject+verb+object) "She takes medicine" {she+takes+medicine}, same sentence in Hindi pattern (subject+object+verb) is उसने दवा ली । (usne dava lee) {उसने+दवा+ली ।}.

A grammatical analysis is needed to produce rules for language translation. Both English and Hindi have their own grammars and we need the proper mapping of English to Hindi grammar. The grammar category for English sentence is "I give him Bryonia and Sulphur" stated as: N+V+(PRON+N1+CONJ+N2). Grammar category for Hindi sentence is "मैंने उसे ब्रायोनिया और सल्फर दी ।" (maine use bryonia aur sulphur di.) stated as: N+V+(PRON+N1+CONJ+N2)

Whereas, N-Head Noun, V-Verb, PRON-Pronoun, N1-Noun, CONJ-Conjunction, N2-Noun.

A stemming and PoS tagging is done while mapping from English to Hindi grammar. It is vital to compare the grammar structure of two languages for proper translation. Rules can be comprehended on the constraint basis, The related works for translation rules in machine translation for Indian languages are discussed in [11, 12, 20]. Assembly rules are classified according to different categories of sentences (simple sentence, compound sentence and interrogative sentence), we designed a total of 163 rules whereas, 45 rules are designed for simple sentence, 53 rules for compound sentence and 65 rules for interrogative sentence. Some of the rules are discussed in Tables 2, 3 and 4 with examples.

Table 2. Assembly rules for simple sentences.

English Rule No.	English Pattern Rules (Er)	Hindi Rule No.	Hindi Pattern Rules (Hr)
Er1 Ex.	Subject | verb(s)| object Aconite is chosen.	Hr1 Ex.	Subject | object | verb(s) एकोनाइट चुना जाता है । (*aconite chuna jata hai*)
Er2 Ex.	Subject | verb(s)| indirect object | direct object | place | time I will give you the report at the hospital tomorrow	Hr2 Ex.	Subject | indirect object | direct object | time | place | verb(s) मैं आपको रिपोर्ट कल अस्पताल में दूंगा । (*main aapko report kal aspatal may dunga*)

Table 3. Assembly rules for compound sentences.

English Rule No.	English Pattern Rules (Er)	Hindi Rule No.	Hindi Pattern Rules (Hr)
Er1 Ex.	Interrogative | auxiliary verb| subject | other verb(s) | indirect object | direct object | place |time What would you like to tell me?	Hr1 Ex.	Subject | indirect object | Interrogative | other verbs | auxiliary verb | direct object | place |time आप मुझसे क्या कहना चाहते हो ? (*aap mujse kyaa kahna chahte ho*)
Er2 Ex.	Interrogative | verb(s)| object Who told you?	Hr2 Ex.	object | Interrogative | verb(s) तुमसे किसने कहा ? (*tumse kisne kaha*)

Table 4. Assembly rules for interrogative sentences.

English Rule No.	English Pattern Rules (Er)	Hindi Rule No.	Hindi Pattern Rules (Hr)
Er1 Ex.	Subject | verb(s)| indirect object | direct object | place |time| Conj.| Subject | verb(s)| indirect object | direct object | place |time I will examine my patient at the clinic tomorrow because I don't have time now.	Hr1 Ex.	Subject | time | place | indirect object | direct object | verb(s)| Conj.| time | Subject | indirect object | direct object | place | verb(s) मैं कल क्लीनिक में अपने रोगी की जांच करूंगा क्योंकि अभी मेरे पास समय नहीं है । (*may kal clinic may apne roogi ki jaach karunga kyoki aabhi mere pass samay nahi hai*)
Er2 Ex.	Subject | verb(s)| indirect object | direct object |Number | time I gave you Bryonia 200 yesterday.	Hr2 Ex.	Subject | time | indirect object | direct object |Number| verb(s) मैंने कल आपको ब्रायोनिया 200 दिया था । (*maine kal aapko bryonia 200 diya tha*)
Er2 Ex.	Time |Subject | verb(s)| indirect object | direct object| Number Yesterday I gave you Bryonia 200.	Hr2 Ex	Time |Subject | indirect object | direct object| Number| verb(s) कल मैंने आपको ब्रायोनिया 200 दिया था । (*kal maine aapko bryonia 200 diya tha*)

Whereas, English Rule 1 (Er1), English Rule 2 (Er2), … and so on. Hindi Rule 1 (Hr1), Hindi Rule 2 (Hr2), … and so on.

Tables 2, 3 and 4 describe the assembly rules of interrogative sentences, simple sentences and compound sentences respectively.

The graph shown in Figure 4 describes the performance of assembly rules, with different percentages.

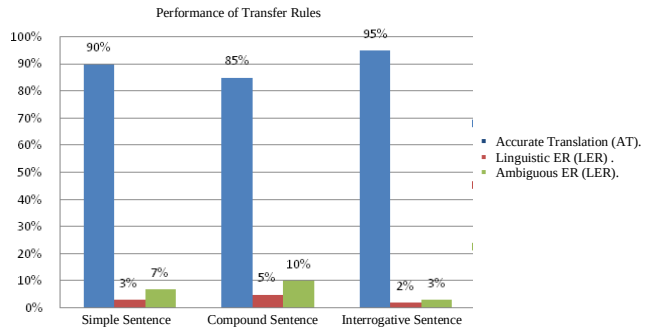


Figure 4. Accuracy graph of assembly rules.

7. Result and Discussion

In Tables 2, 3 and 4, a few assembly rules have been discussed for simple, compound and interrogative sentences respectively along with the examples homoeopathy sentences. To test the precision of the rules, we took 500 sentences from various sources such as Homoeopathy literatures, sentences related to symptoms and articles, as it affects the entire rule based translation process. These sentences have been grouped in simple, compound and interrogative sentence. In the proposed approach, we obtained 90% of correctness for simple sentences, 85% of accuracy for compound sentences and 95% of accuracy for interrogative sentences. The performance of assembly rules for all 500 sentences is measured, for simple sentences; we achieved 90% accuracy, 3% Linguistic Error Rate (LER) and 7% Ambiguous Error Rate (AER). For compound sentences, 85% accuracy, 5% LER and 10% is AER. For interrogative sentences, 95% accuracy, 2% LER and 3% is AER. The result shows 88.33% as the total accuracy of assembly rules for all types of sentences. Over all we achieved a very good accuracy of the mapping rules and we expect the translation accuracy of the system to be high.

8. Conclusions

In the present work, attempts have been made to deal with the translation rules of homoeopathic language with respect to their sentence structure. By going through the literature related to MT, we observed that MT systems often fail to deliver good accuracy for specific domains. This is true even for many open domain systems including google translator. The reason being the lack of relevant corpus and domain specific mapping rules. Therefore, Homoeopathy corpuses for English and Hindi of sizes 20085 and 20072 words respectively have been developed. We use 500 sentences of training data for calculating the performance of assembly rules on the basis of sentence categories like simple, compound and Interrogative sentences. Hence, we conclude simple sentences as: 90% accuracy, 3% LER and 7% AER. For compound sentences we achieved 85% accuracy, 5% LER and 10% AER. For interrogative sentences as 95% accuracy, 2% LER and 3% AER. The result shows 88.33% total accuracy of assembly rules.

References

[1] Al-Taani A. and Abu Al-Rub S., "A Rule- Based Approach for Tagging Non-Vocalized Arabic Words," *the International Arab Journal of Information Technology*, vol. 6, no. 3, pp. 320-328, 2009.

[2] Bahadur P., Jain A., and Chauhan D., "EtranS-A Complete Framework for English to Sanskrit Machine Translation," *the International Journal of Advanced Computer Science and Applications*, vol. 2, no. 1, pp. 52-59, 2012.

[3] Batra K. and Lehal G., "Rule Based Machine Translation of Noun Phrases from Punjabi to English," *the International Journal of Computer Science Issues*, vol. 7, no. 5, pp. 409-413, 2010.

[4] CLAWS Part-of-Speech Tagger for English., available at: http://ucrel.lancs.ac.uk/claws/, last visited 2013.

[5] CliniTrans: Professional Medical Translation Services., available at: http://www.1-800-translate.com/CliniTrans, Last visited 2013.

[6] Dwivedi S. and Sukhadeve P., "Rule based Part of Speech Tagger for Homoeopathy Clinical Realm," *the International Journal of Computer Science*, vol. 8, no. 2, pp. 350-354, 2011.

[7] Francisca J., Mamun M., and Rahman M., "Adapting Rule Based Machine Translation From English to Bangla," *Indian Journal of Computer Science and Engineering*, vol. 2, no. 3, pp. 334-342, 2011.

[8] Jurafsky D. and Martin J., *Speech and Language Processing*, Prentice-Hall, 2000.

[9] Jurafsky D. and Martin J., *Speech and Language Processing an Introduction to Natural Processing Computational Linguistics and Speech Recognition*, Prentice-Hall, 2002.

[10] Krovetz R., "Viewing Morphology as an Inference Process," *in Proceedings of the 16th International Conference on Research and Development in Information Retrieval*, PA, USA, pp. 191-202, 1993.

[11] Rahul C., Dinunath K., Ravindran R., and Soman K., "Rule Based Reordering and Morphological Processing for English-Malayalam Statistical Machine Translation," *in Proceedings of International Conference on Advances in Computing, Control and Telecommunication Technologies*, Kerala, India, pp. 458-460, 2009.

[12] Raji P., "Reordering Approach in English-Malayalam Statistical Machine Translation," *Master's Thesis*, Coimbatore, India, 2010.
[13] Shabdkosh 'kCnd¨'k: English Hindi Dictionary and Translation., available at: http://www.Shabdkosh.com, last visited 2013.
[14] Sinha R., Sivaraman K., Agrawal A., Jain R., Srivastava R., and Jain A., "ANGLABHARTI: A Multilingual Machine Aided Translation Project on Translation from English to Indian Languages," *in Proceedings of International Conference on Systems, Man and Cybernetics Intelligent Systems for the 21st Century*, Vancouver, Canada, pp. 1609-1614, 1995.
[15] Sukhadeve P. and Dwivedi S., "Advancement of Clinical Stemmer," available at: http://languageinindia.com/may2011/kommaluricomplete.pdf#page=51, last visited 2013.
[16] Sukhadeve P. and Dwivedi S., "Developing Hindi POS Tagger for homoeopathy Clinical language," *in Proceedings of the 2nd International Conference Advances in Computer Science and Information Technology*, Bangalore, India, pp. 310-316, 2012.
[17] Sukhadeve P. and Dwivedi S., "Enlargement of Clinical Stemmer in Hindi Language of Homoeopathy Province," *in Proceedings of the 2nd International Conference Advances in Computer Science and Information Technology*, Bangalore, India, pp. 239-248, 2012.
[18] The Stanford Natural Language Processing Group., available at: http://nlp.stanford.edu/software/tagger.html, last visited 2013.
[19] Twitter Part-of-Speech Tagging., available at: http ://www.ark.cs.cmu.edu/TweetNLP/, last visited 2013.
[20] Unnikrishnan P., Antony P., and Soman K., "A Novel Approach for English to South Dravidian Language Statistical Machine Translation System," *the International Journal on Computer Science and Engineering*, vol. 2, no. 8, pp. 2749-2759, 2010.

Sanjay Dwivedi obtained his PhD degree from Banasthali Vidyapeeth in the year 2006. He has completed his PhD in the area of web mining. His research interest are web content mining, semantic web, search engine performance evaluation, e-governance etc. He has published many of the valuable research papers in various National and International Journals. He is presently working as Associate Professor of Computer Science Departement, of BBAU, India.

Pramod Sukhadeve obtained his MSc degree in the year 2006 from Nagpur University. His research interest is natural language processing, machine translation system and in homoeopathy. He has published some of the research papers in refereed Journals and International Conferences. Presently pursuing full time research from BBA University (A Central University) Lucknow.