ISSN 2047-3338

Mapping Parallel English-Hindi Sentences Using English-Hindi Dictionary

Shweta Dubey and Vivek Dubey
CSE Department, Shri Shankaracharya Techniqual Campus, Bhilai (C.G), India
dubeyshweta84@gmail.com, vivekdubey22@gmail.com

***Abstract*– In this paper, we present a methodology for one to one (1:1) mapping of parallel English-Hindi parallel sentences. This methodology is based on the development of parallel English-Hindi word dictionary after syntactically and semantically analysis of the English-Hindi source text. We are using this methodology for the English and Hindi sentences, but the methodology can also be used for other languages. As big parallel corpus of English-Hindi pair language is not usually available, we design and develop two strategies to overcome this problem: normalization of tagged English sentences and Hindi sentences, on the one hand; mapping English-Hindi sentence using parallel English-Hindi word dictionary, on the other. Fortunately, this task, word alignment is well known, and some aligning algorithms are freely available.**

***Index Terms*– Normalization, Tagging, Local Word Grouping, Word Mapping, Part of Speech Tagging (POST) and Word Dictionary**

I. INTRODUCTION

MAPPED English-Hindi parallel corpus is centered theme in English-Hindi example based machine translation (EBMT). EBMT systems are very useful in translating same sentences, and so often used in domain-dependent translation, for example translating user manuals [2]. Trained data of system is prepared by dictionary approach (training data) for collecting rules to local word grouping in English and Hindi sentences [1]. This provides mapping of English-Hindi parallel sentences using dictionary. English and Hindi languages are of different sentence structure. The sentence structure of English is Subject-Verb-Object (SVO) and the sentence structure of Hindi is Subject-Object-Verb (SOV). Many times the number of words in parallel English-Hindi is not same. Before mapping such sentences of English-Hindi are processed to normalize. There are also ambiguities in mapping. Many English words are of different Hindi meaning. Mapping is performed among every character into English and Hindi alphabets [3]. This paper describes one to one mapping of English-Hindi sentences.

Proposed system is of two stages. Initially inputs English-Hindi sentences are normalized and later mapping is carried out. English-Hindi sentences of proposed system have been trained on training data to get normalized sentences using English-Hindi multi-words expression. Normalized English-Hindi sentences have been mapped using English-Hindi Dictionary in proposed system.

English words are normally in dictionary kept with part of speech (POS). Parts of speech tagging is essential to syntactic parsing. Syntactic parsing is analysis of text or sentence to learn sequence of words called tokens, and to decide its grammatical structure with available grammatical rules [4].

In rest of paper, section 2 will explain flowchart of proposed system, section 3 will illustrate tagging of English sentences, section 4 will brief normalizations of English-Hindi Sentences, section 5 will explain mapping of normalized English-Hindi Sentences and section 6 will conclude the paper.

II. FLOW CHAR OF PROPOSED SYSTEM

In proposed system, initially parallel English-Hindi Sentences are saved in input file. Hindi sentences are saved in UTF-8 encoding format. Secondly, English sentence has been tagged using software GENIA Tagger. Thirdly, English-Hindi sentences have been normalized using English-Hindi MWE dictionary. Fourthly, English-Hindi sentences have been mapped using English-Hindi dictionary. Lastly, mapping score of English-Hindi sentences has been saved in output file.

The complete thought of flow chart has been shown in Fig. 1. The first process of flow chart is input of English-Hindi parallel sentences. Secondly, tagging of sentences is done using part of speech tagging (POST). Normalization is based on using English-Hindi MWEs. Tagging and normalization are related to training part of proposed system. Finally output of one to one mapping is obtained using English-Hindi dictionary and saved with gained mapping score.

III. TAGGING ENGLISH SENTENCES

Parts-of-speech tagging (POS tagging or POST) is the process of marking up the words in a text (corpus) as corresponding to a particular POS, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

GENIA tagger is free to download from internet. It has been used in proposed EBMT system for tagging input sentences. It is used to POS tagging, tokenization and etc [5].

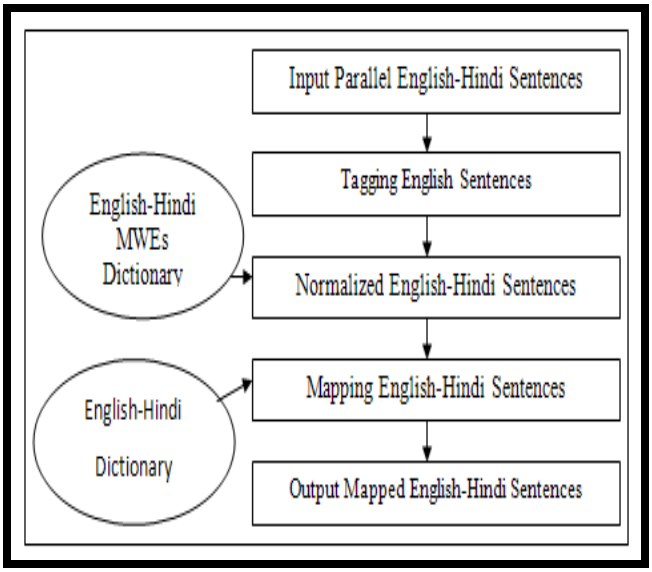

Fig. 1: Flowchart of Proposed System

Table 1: Input-Output of GENIA Tagger

Input Sentence	Output Tagged Sentence
He eats	He/PRP eats/VBZ
He is eating	He/PRP is/VBZ eating/VBG
They did a lot of work	They/PRP did/VBD a/DT lot/NN of/IN work/NN
He is a student	He/PRP is/VBZ a/DT student/NN

Table 2: English-Hindi Multiword Expression

English Sentences	Normalized English Sentences
This/DT girl/NN does/VBZ **no/DT work/NN** on/IN Sunday/NNP	This/DT girl/NN does/VBZ **no_ work/DT** on/IN Sunday/NNP
He/PRP did/VBD all/PDT his/PRP$ work/NN **in/IN time/NN**	He/PRP did/VBD all/PDT his/PRP$ work/NN **in_ time/IN**
They/PRP did/VBD **a/DT lot/NN of/IN work/NN**	They/PRP did/VBD **a_ lot_ of_ work/DT**
We/PRP do/VBP our/PRP$ work/NN **very/RB early/RB in/IN the/DT morning/NN**	We/PRP do/VBP our/PRP$ work/NN **very_ early_ in_ the_ morning/RB**

The input-output of tagging English sentences have shown in Table1. It takes input of English sentences and gives output of each sentence with POS Tagging (POST).

IV. NORMALIZATION OF ENGLISH-HINDI SENTENCES

Normalization of English-Hindi sentences is needed to detect multi word expressions (MWEs) and to remove ambiguity to number of words in English-Hindi parallel sentences. Normalization is performed using English-Hindi multi-words expression dictionary.

A multiword expression (MWE) is understood as out of word boundaries and used underscore "_" in words sequence. For example no_ work is used in single word. Understanding of the word sequence is prepared in one complete word [6].

MWEs act like words and Phrases as based on their construction style. Accurate use of MWEs is useful for different applications like information retrieval, building ontologies, text alignment, and machine translation [7].

MWEs are written with dashes instead of inter-token spaces due to their different structure. Mix methods that merge words as statistically with linguistic information, use morphological, syntactic and semantic ideology to extract MWEs.

Syntactic variety of MWEs is related to different part of speech categories. Different combination of words in structure of MWEs gives tough job of detecting MWEs [8]. Samples of English MWE and Hindi MWE have been shown in Table 2.

MWEs have been constructed using local word grouping where underscore is used in each space of word. For example: *no_work* is one MWE of English sentence. Each MWE is work like one token of sentence, which is easy to map in one to one order. So one to one mapping has been become easier.

T able 3: Normalization Process of English Sentences

English MWE	Hindi MWE
no_ work	कोई_काम_नहीं
in_ time	समय_के_अंदर
a_ lot_ of_ work	बहुत _काम
very_ early_ in_ the_ morning	सुबह_बहुत_जल्दी

Table 4: Normalization Process of Hindi Sentences

Hindi Sentence	Normalized Hindi Sentences
वह अपना काम **नियमित रूप से करता है**	वह अपना काम **नियमित_रूप_से करता_है**
यह लड़की **कोई काम नहीं** इतवार को **करती है**	यह लड़की **कोई_काम_नहीं** इतवार को **करती_है**
उसने सारा अपना काम **समय के अंदर कर लिया**	उसने सारा अपना काम **समय_के_अंदर कर_लिया**
हम **सुबह बहुत जल्दी** अपना काम **करते हैं**	हम **सुबह_बहुत_जल्दी** अपना काम **करते_हैं**

MWEs are used to express expression in effective way. MWEs give same number of words into English-Hindi sentences and make normalized English-Hindi sentences. Normalized English-Hindi sentences make easier the one to one mapping of English-Hindi sentences. Sample of English sentence normalization and Hindi sentence normalization has been explained in Table 3 and Table 4 respectively.

V. MAPPING ENGLISH-HINDI SENTENCES

After normalization of English-Hindi sentences, one to one (1:1) mapping has been carried out using English-Hindi dictionary. English-Hindi dictionary has been described in Table5. Saved input English-Hindi sentences in C++ are shown in figure 2 and figure 3, Dictionary of English-Hindi words using tagging and MWEs in C++ are shown in figure 4 and figure 5. Finally Mapped English-Hindi sentences have been illustrated in figure 6, implemented in C++.

Trained English-Hindi sentences are easier to map because these sentences are contain same meaning, sequence and grammar as shown in Table 5. Final output is show one to one mapping as shown in figure 6, implemented in C++.

One to one mapping has exact one word to one meaning in English-Hindi parallel sentence. Numeric values in result of one to one (1:1) mapping show mapping score of English-Hindi words related to sentence and dictionary, which are same to dictionary and different to sentence as shown in Fig. 6.

For example “is” English word has score E: 2: 1: H: 1:15 one to one mapping to Hindi word i.e., “is” has fifth place in English sentence and fifteen index in English dictionary. Similarly fourth place of Hindi word in Hindi sentence and fifteen index in Hindi dictionary.

Table 5: English-Hindi word dictionary

POSed English Word	Hindi Word	POSed English Word	Hindi Word
a/DT	एक	girl/NN	लड़की
beautiful/JJ	सुन्दर	good/JJ	अच्छा
beggar/NN	भिखारी	He/PRP	वह
blind/JJ	अन्धा	is/VBZ	है
boy/NN	लड़का	man/NN	आदमी

Fig. 2: Shows Input English Sentences in C++

Fig. 3: Shows Input Hindi Sentences in C++

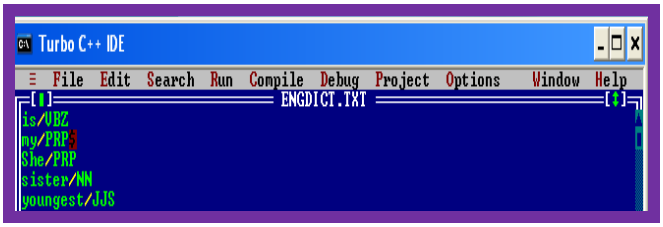

Fig. 4: Shows English dictionary in C++

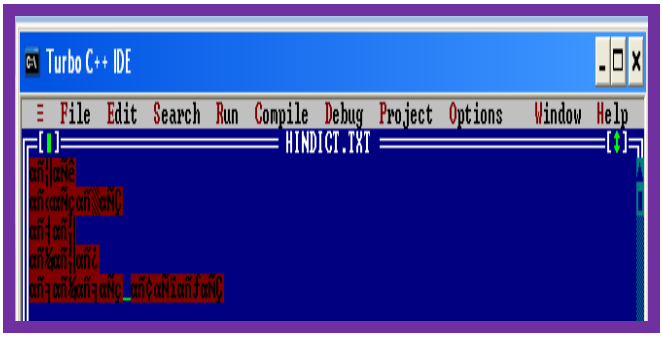

Fig. 5: Shows Hindi dictionary in C++

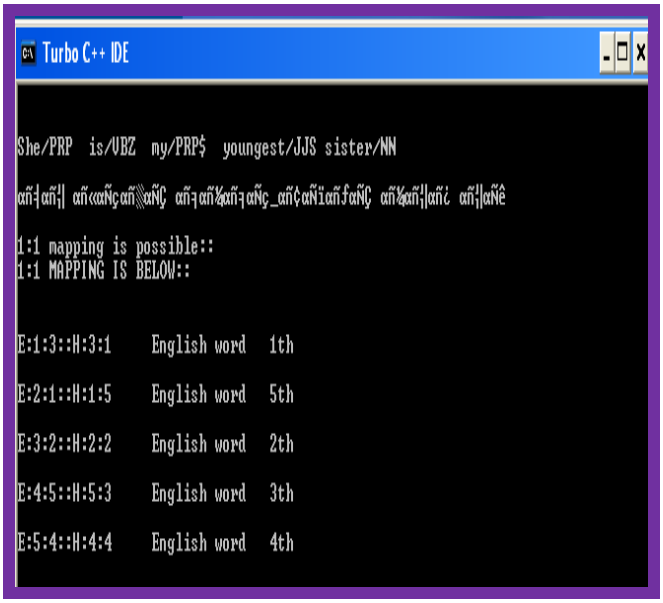

Fig. 6: Shows Score of One to One (1:1) Mapping in C++

VI. CONCLUSIONS AND FUTURE WORK

Proposed one to one (1:1) mapping of English-Hindi sentences methodology is novel and efficient. Experiments on 500 different parallel English-Hindi sentences have been carried out. Result is drastic and motivated to design for different pair of language.

Study of more and other structure, related to English-Hindi word has been left to future work. Also the future works related to displaying Hindi fonts and POST in Hindi sentences in implementation have been left to future work.

REFERENCES

[1] Niraj Aswani, "Aligning words in English-Hindi parallel corpora", Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 115–118.

[2] Tong Xiao, Huizhen Wang, "The NiuTrans Machine Translation System for NTCIR-9 Patent", Proceedings of NTCIR-9, December 6-9, 2011, Tokyo, Japan, Pages 593-599.

[3] Niraj Aswani, "A hybrid approach to align sentences and words in English-Hindi parallel corpora", Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 57–64.

[4] Antony P J, Nandini. J. Warrier, Soman K. P., "Penn Treebank-Based Syntactic Parser for South Dravidian Languages using a Machine Learning Approach", International Journal of Computer Applications (0975–8887), Vol. 7, No. 8, October 2010, pages 14-21.

[5] Yoshinobu Kano, Jun'ichi Tsujii, "Sharable Type System Design for Tool Inter-Operability and Combinatorial Comparison", The First International Conference on Global Interoperability for Language Resources, pages 121-129.

[6] Richard Beaufort, Sophie Roekhaut, Louise-Amélie, Cougnon Cédrick Fairon, "A hybrid rule/model-based finite-state framework for normalizing SMS messages", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 770–779.

[7] Hassan Al-Haj, Shuly Wintner, "Identifying Multi-word Expressions by Leveraging Morphological and Syntactic Idiosyncrasy", Proceedings of the 23rd International conference on Computational Linguistics, 2010, pages 10–18.

[8] Yulia Tsvetkov, Shuly -Wintner, "Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources", Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 836–845.

Shweta Dubey has received her B.Sc. and M.Sc. degree in Computer Science from SSMV (Shri Shankaracharya Mahavidyalay) in Bhilai by Pt. Ravishankar Shukla University, Raipur (C.G.), at 2005 and 2007 respectively. Currently, pursuing the Master of Engineering (M.E) Degree in Computer Technology and Application (C.T.A) from Shri Shankaracharya Group of Institution (SSGI), Chhattisgarh Swami Vivekananda Technical University (CSVTU), Bhilai (Chhattisgarh).

Vivek Dubey is the Associate Professor in the CSE Department of Faculty of Engg. & Tech, Shri Shankaracharya Group of Institute Bhilai, CG, India. He is the Incharge of NLP and Speech Laboratory at the Institute. He did BE (CSE), M.Tech (CT) and pursuing his Ph.D. in Computer Science & Engineering. He has 13 years of Engg. teaching experience and 7 years in other. He has published around 35 papers in various national and international journals/conferences. He is the Editor and Reviewer in various journals.