Supporting Large English-Hindi Parallel Corpus using Word Alignment

Shweta Dubey
Assistant Professor
Bhilai School of Engineering
Bhilai, India

Tarun Dhar Diwan
Assistant Professor
Dr. CV Raman University
Bilaspur, India

ABSTRACT
This paper gives description about methodology to understand parallel English-Hindi sentences using word alignment. This methodology is foundation to develop the parallel English-Hindi word dictionary after syntactically and semantically analysis of the English-Hindi source text. Methodology of proposed system is used for the English and Hindi sentences; also the methodology can be used for other languages. Outsized parallel corpus of English-Hindi pair language is not frequently available. Development is based on two strategies to solve this problem. First is normalization of tagged English sentences and Hindi sentences. Second is mapping English-Hindi sentence using parallel English-Hindi word dictionary. Fortunately word alignment is clearly known and few aligning algorithms are without restraint accessible.

Keywords
Tagging, Local Word Grouping, Word Mapping, Normalization, Part of Speech tagging(POST), Word Dictionary, Multi Word Expressions, Mapping Score.

1. INTRODUCTION
Over all main idea of mapped English-Hindi parallel sentences is in English-Hindi example based machine translation (EBMT). EBMT systems are very helpful in translating same sentences, and so often used in domain-dependent translation, for example translating user manuals [2].

Dictionary approach (training data) prepares trained data of system for collecting rules to local word grouping in English and Hindi sentences [1]. Dictionary provides easiest mapping of English-Hindi parallel sentences using dictionary. English and Hindi languages have different sentence structure. Sentence structure of English is Subject-Verb-Object (SVO) and the sentence structure of Hindi is Subject-Object-Verb (SOV).

A lot of times the number of words in parallel English-Hindi is not same. Such sentences of English-Hindi have to normalize before mapping. Ambiguities are also in mapping. Many English words are of dissimilar Hindi meaning. Mapping is performed along with every character into English and Hindi alphabets [3]. This paper explains one to one mapping of English-Hindi sentences.

Proposed system contains two stages. Firstly inputs English-Hindi sentences are normalized and afterward mapping is carried out. English-Hindi sentences of proposed system have been trained on training data to obtain normalized sentences by English-Hindi multi-words expression. Normalized English-Hindi sentences have been mapped by English-Hindi Dictionary in proposed system.

English words are usually in dictionary kept with part of speech (POS). Parts of speech tagging is necessary to syntactic parsing. Syntactic parsing is study of text or sentence to find out sequence of words called tokens, and to choose its grammatical structure with available grammatical rules [4].

In organization 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.

2. FLOW CHART OF PROPOSED SYSTEM
In proposed system, firstly parallel English-Hindi Sentences are saved in input file. Hindi sentences are saved in UTF-8 encoding format. Secondly, English sentence has been tagged with POS (part of speech) by software GENIA Tagger. Thirdly, English-Hindi sentences have been normalized using English-Hindi MWE dictionary. Fourthly, English-Hindi sentences have been mapped by English-Hindi dictionary. Lastly, mapping score to alignment of English-Hindi sentences has been saved in output file.

The whole thought of flow chart has been shown in Figure 1. The first procedure of flow chart is input of English-Hindi parallel sentences. Secondly, tagging of sentences is prepared 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. Lastly output of aligned and one to one mapping is obtained using English-Hindi dictionary and saved with gained mapping score.

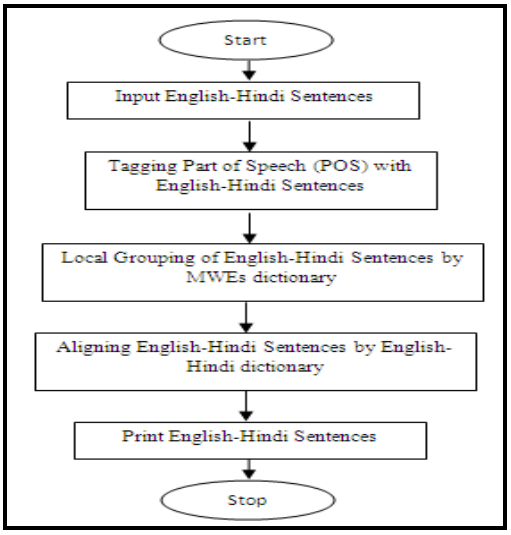


Figure 1: Flowchart of Proposed System

3. TAGGING ENGLISH SENTENCES

Parts-of-speech tagging (POS tagging or POST) is the procedure of marking up the words in a text (corpus) as corresponding to a particular POS, based on both its meaning, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. GENIA tagger is free of charge 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]. The input-output of tagging English sentences have specified in Table1. It takes input of English sentences and gives output of each sentence with POS Tagging (POST).

Table 1: Input-Output of GENIA Tagger

Input Sentence	Output Tagged Sentence
I am eating	I/PRN am/VBZ eating/VBZ
Ramesh is eating	Ramesh/NN is/VBZ eating/VBZ
Girls did a lot of work	Girls/NN did/VBZ a/DT lot/NN of/IN work/NN
Rupal is a student	Rupal/NN is/VBZ a/DT student/NN

4. NORMALIZATION OF ENGLISH-HINDI SENTENCES

Normalization of English-Hindi sentences is desirable to align and 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 known as out of word boundaries and used underscore "_" in words sequence. For example no_ work is used in single **no_work** is one MWE of English sentence. Each MWE reacts like one token of sentence, which is easy to map in one to one order. Tagging of POS in MWEs follows the POS of last word in MWE. E.g. **no/DT work/NN-> no_ work/NN**. Finally, one to one mapping has been become easier due to MWE.

word. Understanding of the word sequence is prepared in one complete word [6].

MWEs proceed like words and Phrases as based on their construction style. Exact use of MWEs is valuable for different applications like information retrieval, building ontology, text alignment, and machine translation [7].

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

Syntactic variety of MWEs is associated 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 specified in Table 2.

Table 2: English-Hindi Multiword Expression

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

MWEs have been created using local word grouping where underscore is used in each space of word. For example:

T able 3: Normalization Process of English-Hindi Sentences

English Sentences	Hindi Sentences	Normalized English Sentences	Normalized Hindi Sentences
I/PRN did/VBZ a/DT lot/NN of/PRP work/NN	मैं बहुत काम किया	I/PRN did/VBZ a_lot_of_work/NN	मैं बहुत _काम किया

MWEs are used to utter expression in effective way. MWEs provide same number of words into English-Hindi sentences and create normalized English-Hindi sentences.

Normalized English-Hindi sentences create easier alignment using one to one mapping of English-Hindi sentences. Sample of English sentence normalization and Hindi sentence normalization has been described in Table 3.

5. MAPPING ENGLISH-HINDI SENTENCES

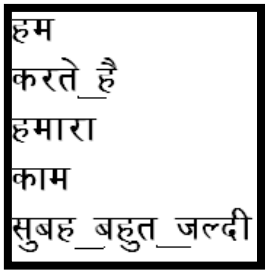
After normalization of English-Hindi sentences, alignment using one to one (1:1) mapping has been performed using English-Hindi dictionary. English-Hindi dictionary has been explained in Table5. Saved input English-Hindi sentences are specified in figure 2 and figure 3, Dictionary of English-Hindi words using tagging and MWEs are specified in figure 4 and figure 5. Finally aligned and mapped English-Hindi sentences have been explained in figure 6, implemented.

Trained English-Hindi sentences are very easy to map because these sentences are contain same meaning, sequence and grammar as specified in Table 5. Final output is explained one to one mapping as shown in figure 6, implemented.

Table 4: English-Hindi word dictionary

POSed English Word	Hindi word	POSed English Word	Hindi word
She/PRN	यह	boy/NN	लड़का
Work/NN	काम	girl/NN	लड़की
Morning/NN	सुबह	Sunday/NNP	इतवार

Figure 4: Shows English dictionary



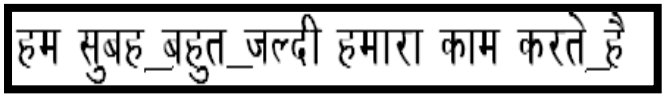


Figure 2: Shows Input English Sentences

Figure 3: Shows Input Hindi Sentences

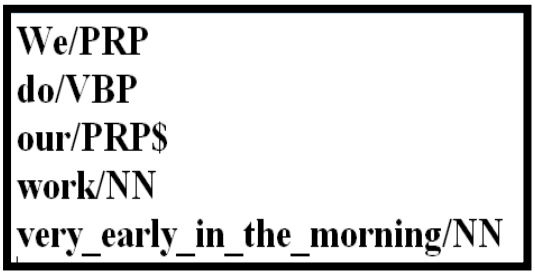
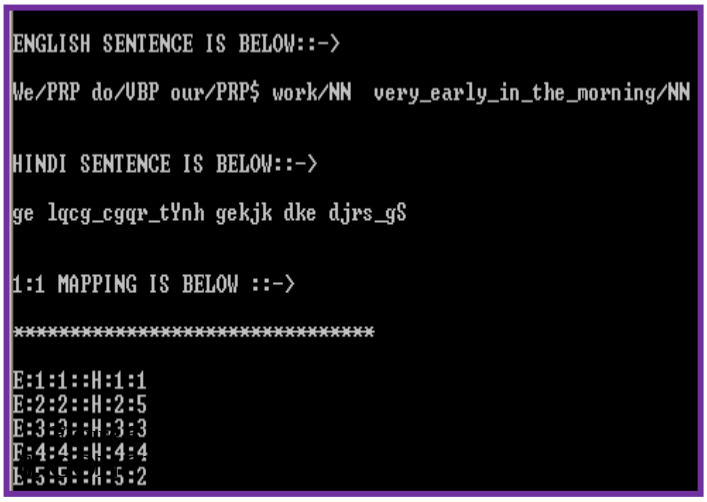


Figure 5: Shows Hindi dictionary

One to one mapping contains accurate one word to one meaning in English-Hindi parallel sentence to support full of knowledge based alignment. Numeric values in result of one to one (1:1) mapping explain mapping score of English-Hindi words related to sentence and dictionary, which are similar to dictionary and different to sentence as shown in figure 6.

For example English-Hindi sentence which is shown by figure 2 and figure 3, gives mapping score as shown in below figure 4 with proper alignment.



Proposed one to one (1:1) mapping based alignment of English-Hindi sentences is very helpful to understand meaning of every expression and word from English to Hindi or Hindi to English. Experiments on 555 different parallel English-Hindi sentences have been accepted. Result is well-built and enforced to design for different pair of language.
Future work is left to study about more other type of alignment to English, Hindi and other languages.

7. REFERENCES

[1] Niraj Aswani, "Aligning words in English-Hindi parallel corpora", Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages115–118.

[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, Dr. Soman K P, "Penn Treebank-Based Syntactic Parsers for South Dravidian Languages using a Machine Learning Approach", International Journal of Computer Applications (0975 – 8887), Volume 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 AnnualMeeting 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, pages 10–18.

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