International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S, August 2019

Issues in Word Alignment from Hindi-English Languages

Shachi Mall, Umesh Chandra Jaiswal

ABSTRACT--- The paper discusses the various methods and issues related to word alignment. This paper focus on the main problem arises in word alignment because Hindi language is based on subject object verb "SVO" and for English language is subject verb objects "SOV". Hindi is morphology rich language, therefore correct alignment of word order from Hindi to English language is quite difficult. The paper presents survey on for foreign and Indian language of word alignment in the application of machine translation

Keywords— Word alignment; semi supervised; unsupervised; machine translation.

I. INTRODUCTION

Word alignment is done in machine translation to identify the correct sense and align the given input Hindi language into target language means in English language. The problem of word alignment for Hindi to English language is word order Hindi into English language. To identify the each word by relationship in the pair of sentences is obtained in word alignment which should be correctly aligned [1]. Figure 1 illustrates the process of word alignment.

For example consider Hindi sentence -मुझे इस पुस्तक के पहले पृष्ठ की एक प्रति चाहिए।

English translation- I need a copy of the first page of this book.

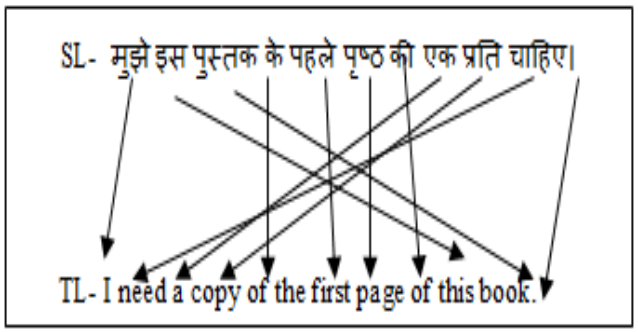

Figure 1: Example of word alignment from Hindi to English language

Manuscript published on 30 August 2019.
***** Correspondence Author (s)
Shachi Mall, Department of Computer Science & Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur, U.P, India. (Email: shachimall@gmail.com)
Umesh Chandra Jaiswal, Department of Computer Science & Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur, U.P, India(Email: ucjjaiswal@yahoo.co.in)

There are various methods for word alignment which are classified in unsupervised and supervised approaches. These approaches are also used in the combination to improve the quality of word alignment. During literature survey we find few research work has been carried in word alignment for different languages. To develop word alignment for Hindi to English language is a challenging task.

Supervised based approach for word alignment for Hindi to English languages are time consuming task it needs sense annotated corpus for the Hindi language [2].

Unsupervised approach uses WordNet [3] this method has the potential to resolve the drawback of large scale corpus has been manually annotated for word alignment. This approach is divided into three different categories: Graph based, word clustering and context clustering. Vector space model is popular approach used in graph based approach [4].

Another popular method word alignment based on Statistical based machine translation [5]. This is different method based on global lexical selection method.

IBM model uses GIZA++ tool for word alignment [6]. In which it automatically correlate with bilingual lexical.

The paper compares different approaches used for word alignment and find what the problems arises in different approaches.

The paper is organized in following sections: The related work was discussed in section 2. In section 3 different approaches for word alignment and 4 we discuss the objectives of the research work. In section 5 Conclusion.

II. RESULTS & DISCUSSIONS

A number of Indian researchers have carried out their work related to machine translation for Indian languages. Less work has been carried out in Hindi language. In Hindi language word alignment is an important task to produce the correct translation from Hindi to English Machine Translation. Since their in-troduction in the work of Dhariya et. al. [7] is the conversion of Hindi-English language combined Rule based, Example based and Statistical based approach to enhance the quality of translation was improved due to morphological analyzer. Identify the semantic similarity in the complex Hindi sentences WordNet is used to detect all the semantic relations [8]. Dependency parsing is another approach in which the dependency parse tree was is created for source language. Many supervised and unsupervised based method is used for Word Sense Disambiguation to improve the quality of translation [9]. Word alignment is another issue in Hindi to English

language translation many approaches such as IBM model and Expectation-Maximization (EM) algorithm.

A. *Foreign Languages*

Word alignment application was used for different foreign languages machine translation but the results are quite promising. Statistical based word alignment method to translate French language in to English language [11]. The approach used IBM model-1 which automatically extracts

the vocabulary the shortcoming of the model was of large vocabulary due to this it not align all the French word as source language into English as a target language. The another research was done using dynamic programming approach in which word by word translation was done by using contingency table that provide paired word information, the system accuracy was 60% of word alignment [12]. Another research was carry using unsupervised based vector space model [13]. This model is based on context clustering in which co-occurrence method is used. The source language words are tokenized from the sentence and each word assign as vector in co-occurrence. All words are clustered in group and each group as identified as sense of target word and correct alignment. Graph based word alignment in which it uses grammatical relations between the context words [book]. In this method they developed Markov clustering algorithm to identify the word alignment. Another attempt in graph based method was agglomerative cluster in which edges are constructed from corpus of the co-occurrence graph [14].

B. *Indian Languages*

The word alignment done for Hindi to English language based IBM model the author try to resolve the limitation of IBM model by improving Part of speech tagging method which reduces the execution time and enhance the performance [15]. The author test the system accuracy by taking dataset of 270 corpus and compare the developed system by incorporating the part of speech tagging method in IBM model-1 and compare with IBM model-1. The F-measure accuracy result of the developed system was 56.42% and the F- measure accuracy result of IBM model-1 was 42.94% and Alignment Error Rate (AER). result accuracy of IBM model-1 is 57.06% and accuracy result of IBM model with incorporating part of speech tag was 43.58%. the drawback of the IBM model is many translation for one word, in long sentence all words are not align and some hindi word have multiple translation.

Statistical based Marathi to Hindi word alignment [Marathi to hindi]. In this method Marathi words are splitted for this they developed devised algorithm to split compound Marathi word and uses list of siifix the list was developed using bilingual and monolingual corpus. Marathi to Hindi word alignment was trained on GIZA++ tool. The accuracy result was test for BLUE, NIST, PER AND WER are respectively 38.35%, 7.756%, 42.08% and 35.82%

English – Hindi word alignment based on corpus augmented resource [16]. This approach uses two tools GIZA++ and NATools. The Giza++ tool for word alignment and NATools was used for bidirectional dictionary. The system take Hindi as source language as input and preprocess, the NATools is used for lexical extraction and the output of this submodule is given input to GIZA++ for word alignment. The absolute reduction of the system was 5.96%.

Word alignment based on global lexical reranking approach [17]. The main advantage of this approach is local association to obtain Hindi to English lan-guage another advantage is global lexical selection in this method the system extract the lexico-syntactic features of words from source language. The result was tested for Mosses and BLEU. The accuracy result of global lexical selection reranked score 2.24that was higher than the Mosses.

C. *Identification of Research Gap and Problem*

There are many unresolved issues word alignment in application of machine translation for Indian languages such as:

a. Morphological analyses faces problem in productivity and creativity in languages, word that are not licensed than it will remain unparsed. This is known as unknown word.

b. Grammatical tagging corpora and Chunk the sentence

c. Construction of electronic dictionary.

d. Resolve the word sense disambiguation. The major drawback is the problem of scale.

e. Word alignment for Hindi to English Languages.

f. Issues related to the long sentences fails in word alignment.

g. Issues related to vocabulary when they are very large.

h. Anusaaraka fails to resolves alignment of the word in English-Hindi translation [5].

i. Expectation-Maximization (EM) algorithm and Gibbs sampling approach was used for word alignment for rare words and morphologically rich for small corpora. The drawback was small corpora size [8].

j. Multi-Model approach based on regression model which dynamically select the sentence for multiple translation. The experimental result was shown on baseline system. The semantic information improves the performance. The drawback was in phrase representation i.e. 'forest'+'black' and 'black'+'forest' shows the same vector representation but in reality they are different [9].

k. Sequential word labeling based reordering rules suggestion which increase the performance of parser but it has following drawback [10RS]

table 1: different approaches, results and drawback of word alignment

S.No.	Author	Approach	Performance	Drawback
1	Sen S et. al. [19]	Statistical Machine Translation	Official baseline BLEU scores of 10.79%	Word sense disambiguation and word alignment was not resolved
2	Nair J et. al. [20]	Statistical and Rule based	-	Translation is fails for complex sentences due to limitation of bilingual dictionary and word alignment was not done
3	Srivastava J et.al. [18]	IBM model, GIZA++ and clause identification	AER 51.96%	Morphology was not resolved and multiple translation for one word of Hindi.
4	Srivastava et. al. [12]	Grammar mapping technique and Rule based	-	Word sense disambiguation was not resolved
5	Srivastava J et.al. [15]	IBM model with incorporating POS Tag	AER 48.95%	In large sentence alignment is fails.

There are many approaches and tools are developed for Hindi to English translation but still there are many limitations in the translation if we work on parser this improve the Word Sense Disambiguation and Word alignment.

D Solution of Research Gap and Problem

- Construct the large electronic dictionary which understood the data structure and directly obtains result. This can be done by lookup operation. Large sentence can break into sub sentences.
- Hindi language is morphologically rich language to overcome with the shortcoming we can incorporate Morphological parsing to identify the constitute morphemes. Hindi uses the suffix and prefix information to express inflectional and derivational morphology. Shallow parsing perform pruning to reduce the Morphological analyzer ambiguity.
- Hindi language is morphologically rich language to overcome with the shortcoming we can incorporate Morphological parsing to identify the constitute morphemes. Hindi uses the suffix and prefix information to express inflectional and derivational morphology. Shallow parsing perform pruning to reduce the Morphological analyzer ambiguity.
- The modified Lesk‟s algorithm make more concrete for polysemy word.
- Shallow parsing and word sense disambiguation can enhance the quality of word alignment, for this we use GIZA++ tool in which we can take all words i.e connecting words article etc. and resolve the issues of separating the connecting words articles.

III. APPROACHES FOR WORD ALIGNMENT

Word alignment is broadly classified in to following categories:

A. Lexical based

Lexical based approach is used for word alignment is classified in two types lexical item and lexical reordering. Lexical item is used to find the lexical item from input language to match the output language and lexical reordering is used to reorder the source word into target word with correct alignment. The combination of both approach also contain syntactic information of the source language. For this we use we use bag of word model. In this model feature extraction was done for source sentence and for each word find the target word translation, the target word are permuted in permutation window and best target word are selected. The drawback of this model is the long distance word reordering was not allow, to overcome with this limitation sequential lexical choice model is used.

B. Statistical based

Statistical based word alignment uses large parallel data between source language and target language. IBM model (1-5) and Hidden Markov model is based on statistical based machine translation. To evaluate the result Arithmetic error rate and performance of IBM model (1-5) and for Hidden Markov model are calculated the formula is as below:

$$AER \approx 1 - \frac{|A \cup T| + |A \cap P|}{|A| + |T|} \qquad (1.1)$$

In equation (1.1) “A” represents alignment, where “T” is sure alignment and “P” possible alignment. Arithmetic error alignment “AER” is calculated by comparing gold standard with IBM model. Hidden Markov model based word alignment in which different algorithm is used i.e. Viterbi, Baum-Welch and the forward-backward algorithm.

IV. OBJECTIVE

Our aims are to design and develop word alignment and used

the application in Hindi to English languages machine transla

-tion in which we resolve the issues and has been discussed in

solution of research gap and problems.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. CONCLUSION

A major drawback in Word alignment in Hindi to English machine translation to overcome with this problem if we use statistical model for sentence alignment with the help of parallel corpus, the translation model works well with the large corpus. The sense word alignment (with probabilities), it can be used for lexicon acquisition for Hindi to English machine translation. Statistical MT techniques have not so far been widely explored for Indian languages.

REFRENCES

1. Dhariya, O., "A hybrid approach for Hindi-English machine translation" In proceeding of IEEE International Conference on In-formation Networking (ICOIN), (pp. 389-394), 2017.
2. Gao, Q" A semi-supervised word alignment algorithm with partial manual alignments", In the proceeding of Association for Computational Linguistics International Conference on the Joint Fifth Workshop on Statistical Ma-chine Translation and Metrics MATR (pp. 1-10).
3. Mihalcea, R. " Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for se-quence data labeling", I In the proceeding of Association for Computational Linguistics International Conference on Hu-man Language Technology and Empirical Methods in Natural Language Processing (pp. 411-418), 2005.
4. Faruqui, M., "Improving vector space word representations using multilingual correlation", In the proceeding of Association for Computational Linguistics International Conference on 14th Conference of the European Chapter (pp. 462-471), 2014.
5. Bangalore, S," Statistical ma-chine translation through global lexical selection and sentence reconstruction", In the proceeding of Association for Computational Linguistics International Conference on 45th Annual Meeting (pp. 152-159), 2007
6. Vogel, S., "Parallel implementations of word alignment tool", In the proceeding of Association for Computational Linguistics International Conference on Software engineering, testing, and quality assurance for natural language processing (pp. 49-57), 2008.
7. Malviya, S., "A hybrid approach for Hindi-English machine translation", In the proceeding of Association for Computational Linguistics International Conference on Information Networking (ICOIN), (pp. 389-394), 2017.
8. Ranathunga, S., "Sinhala Short Sentence Similarity Calculation using Corpus-Based and Knowledge-Based Similarity Measures", In the proceeding of Association for Computational Linguistics International Conference on 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016) (pp. 44-53), 2016.
9. Damani, O. P. ," A domain-restricted, Rule Based, English-Hindi Machine Translation system based on dependency parsing", In the proceeding of Association for Computational Linguistics International Conference on 11th International Conference on Natural Language Processing (pp. 177-185), 2014.
10. Marcu, D. ,"Measuring word alignment quality for statistical machine translation", In the proceeding of Association for Computational Linguistics International Conference on Computational Linguistics, 33(3),(pp- 293-303), 2007.
11. Mercer, R. L. ,"The mathematics of statistical machine translation: Parameter estimation", In the proceeding of Association for Computational Linguistics International Conference on Computational linguistics, 19(2),(pp- 263-311), 1193.
12. Church, K. W. (1991, February), "Identifying Word Correspondences in Parallel Texts", In HLT (Vol.91, pp. 152-157), 1991.
13. A. Y. ," Semantic compositionality through recursive matrix-vector spaces", In the proceeding of Association for Computational Linguistics International Conference on joint conference on empirical methods in natural language processing and computational natural language learning (pp. 1201-1211), 2012
14. Manmatha, R.,"Word spotting for historical documents", In the proceeding of International Journal of Document Analysis and Recognition 9(2-4), (pp-139-152), 2007.
15. Srivastava J., ," Segmenting Long Sentence Pairs to Improve Word Alignment in English-Hindi Parallel Corpora," In the proceeding of International Journal on Advances in Natural Language Processing. Lecture Notes in Computer Science, vol 7614. Springer, Berlin, Heidelberg, 2012.
16. Venkataramani, E,"English-hindi automatic word alignment with scarce resources," In the proceeding of IEEE International Conference on Asian Language Processing (IALP), (pp. 253-256), 2010.
17. Ney, H. ," Extending statistical machine translation with discriminative and trigger-based lexicon models", ," In the proceeding of IEEE International Conference on Asian Language Processing on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 210-218), 2009.
18. Srivastava, J., ," POS-based word alignment for the small corpus", ," In the proceeding of IEEE International Conference on Asian Language Processing (pp. 37-40), 2015.
19. Hambir, N. ," English to Hindi Machine Translator Using GMT and RBMT Approach", ," In the proceeding of IEEE International Conference on Asian Language Processing Advances in Computing and Information Technology (pp. 219-225), 2012.
20. Deetha, R. ," An efficient English to Hindi machine translation system using hybrid mechanism", ," In the proceeding of IEEE International Conference on Asian Language Processing (pp. 2109-2113), 2016.