International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, April 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

# Word Alignment of English-Hindi Parallel Corpus: Relative Study

**Nitika Nigam[1], Umesh Chandra Jaiswal[2]**

[1] *Department of Computer Science and Engineering,*
*Madan Mohan Malviya University of Technology, Gorakhpur, INDIA*
[2] *Department of Computer Science and Engineering,*
*Madan Mohan Malviya University of Technology, Gorakhpur, INDIA*

**ABSTRACT**

*Word alignment is the method of recognizing the words with their correct translated paired language words. The high quality and domain-specific parallel corpus produce a virtuous quality of word alignment and aligned words are useful in extracting the pattern with the semantic meaning. A couple of word arrangement approaches have been used in NLP. Word corpus is applied to get multiword phrases with semantics sense. There have been many methods for word alignment but the work in English-Hindi word alignment is still in development. In this review paper, we cover the techniques utilized as a part of the word arrangement and are quickly portrayed with their features and limitations. Furthermore, it covers the difficulties in the word arrangement of English-Hindi parallel corpus.*

**Keywords-** ***Word alignment, Lexical based approach; Statistical based approach, Natural Language Processing***

## I INTRODUCTION

Word alignment is the part of Natural language processing (NLP), where NLP is an area which provides interactions between computers and human languages. The communication is done with the support of Machine Translation (MT) and MT is the process of translating from one natural language to others. Word alignment is a backbone and an intermediate module for statistical machine translation (SMT) [1].

The word by word relationship in the pair of sentences is obtained in word alignment which should be correctly aligned [1]. The other name of Word alignment is Bi-text word alignment, as the alignment is done between two languages [2]. Several approaches which are used to perform word alignments like a bipartite graph; Vector-based methods etc. are helpful in aligning the words easily [3]. The bipartite graph is a pictorial representation which can contain maximum n^2/4 edges or connection; where n is numbers of words present in the sentences. The main application of word alignment is the automatic extraction of bilingual words and their meaning from the corpora [4]. Word alignments enhance many NLP applications if and only if word alignment is of better quality.

We emphasize on word alignment of English – Hindi data set. In English–Hindi parallel corpora, alignment of words can be one to one, one-to-many or may contain multiple connections [4]. There are many word alignment techniques which have been proposed by many authors in another language. But the work in English- Hindi language is still in development. The reason behind this is that the resources of Hindi languages are very less, and it is morphologically rich language [5]. Alignment between the parallel corpora which can be one to one mapping ( shown in Fig.1) and one-to-many mapping or vice versa is shown in Fig.2.

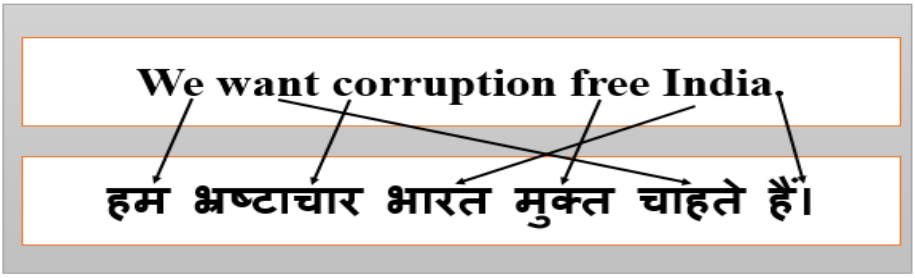


**Fig. 1. The word alignment example of bilingual sentences. (one-one mapping)**

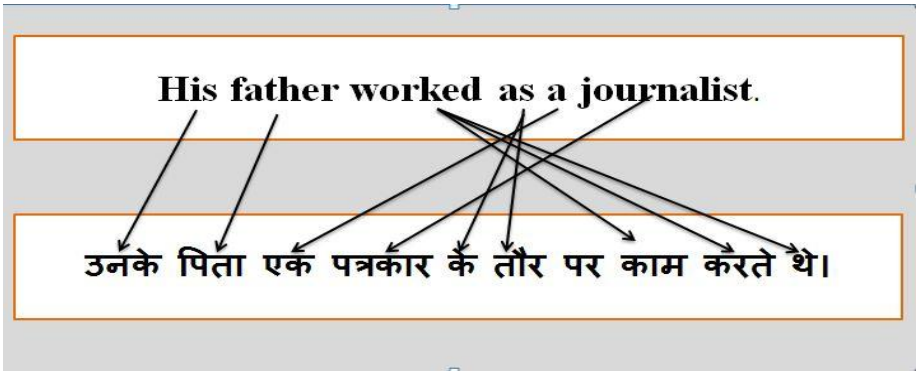


**Fig.2. The word alignment example of bilingual sentences. (one-many mapping and vice versa)**

## II LITERATURE REVIEW

P. F. Brown et al. [1] proposed models for word alignment with the help of parallel corpora for French and English at the sentence level. After sentence level, they classify the word correspondence in the bilingual corpora. They proposed that bilingual lexical correlation can be extracted automatically. They described the IBM models which are a statistical model for translation. The drawback was the IBM model-1 can't handle the large vocabularies and due to which it can't align some words of French with English words. W. A. Gale et al. [6] introduced the method to measure the correspondence of parallel words based on 2X2 contingency table. This provides information of a pair of translated words. This method doesn't provide the good precision of automatic alignment as human expertise is needed after iterations and is treated as noisy data. Thus, dynamic programming techniques were used and 60% of correspondence was achieved. The drawback was that it produces incorrect alignment if huge information set was given.

C. Cherry et al. [7] used the arithmetical based approach in which they improved the performance of alignment at sentence level by computing probability. This model permits easy integration of context-specific features. The drawback of this model is that it only works for one to one alignments only and not for many to one or many to many alignments. Robert C. Moore et al. [8] applied the unpretentious method which improves the precision of IBM model 1. He reduced the Alignment error rate approximately by 30%. The drawback of this system is that it doesn't increase the efficiency of parameter F-measure and recall.

N. Aswani et al. [9] proposed a scheme which aligns the English-Hindi texts both at the word and sentence level. They used simple sentence length approach at the sentence level and hybrid approach at the word level. The dictionary lookup approach was used for multi-features and nearest neighbor approach was for many to many mapping of words. For many-to-many word alignment they gained 99.09% accuracy and 77% precision with 67.79% recall. The drawback of this system doesn't solve the problem of subject-object-verb problem.

Niladri Chatterjee et al. [3] used the recency-vector based approach as it performs better for small data set. They used two algorithms (DK-vec Algorithm and Somers' Algorithm) for the experiments which were on the manual English - Hindi parallel corpus. Since the lexicon based approach is resource dependent thus the recency-vector based approach offers the best substitute. The drawback was that it only works on the small corpus. S. Pal et al [12] proposed the system for extracting the phrases. The hybrid word alignment model was proposed by them for phrase-based statistical machine translation (PB-SMT). The unsupervised tools used were GIZA++ and Berkeley aligner and combined with rule-based word alignment techniques. Berkeley aligner was used to expand the word alignment quality. Amalgamation of these tools helps in extracting the aligned phrases. The drawback of this system was that it doesn't resolve the problem of multiple alignments of bilingual words.

Eknath Venkataramani et al. [4] provided the corpus augmented method of word alignment. They used two GIZA++ tool which performs word alignment statistically and NATools for providing the bilingual dictionary. Hindi is a complex language and the resources are limited, above approach helps in eliminating these limitations. The parallel corpus under goes through 5 stages process and this method help in reducing the Alignment Error Rate (AER) approximately to 5.96%. When the method was applied after the POS tagging, the correct alignments of the adjective class and noun class increases whereas in the verb class the correct alignments slightly increase. Jyoti Srivastava et al. [11] used the approach of divide and conquer; their approach was to divide the large sentences into clauses which are treated as sentence pairs. These pairs were trained on word alignment model. 300 sentences were used for the experiment in which 270 sentences were trained and remaining left is used for testing. Their system increased the F- measure by 10%where decrease the AER (Alignment Error Rate) by 10%.

Jyoti Srivastava et al. [5] used statistical based word alignment model. They trained 950 sentences and 50 sentences were used for testing. In this system, the AER (Alignment Error Rate) was decreased by 4% and F-measure was increased by 4% in comparison of base system for word alignment GIZA++. POS tagger was used

to tag the corpus, which helps in reducing the incorrect alignments. The drawback of the system was that it doesn't work for the large corpus.

## III APPROACHES

Following approaches used in word alignment:

### 3.1 Lexical -based approach

It basically depends upon the lexical resource i.e. the dictionaries. These dictionaries consist of lexical information regarding words, which are classified into nouns, adjective, translation of those words etc. Thus, also known as dictionary based approach.

It can be classified as a dependent approach because it is totally dependent upon the resources provided for alignment process. Stanley F. Chen [13] proposed a method in which he aligned the parallel corpus using lexical information. An algorithm was proposed by I. D Melamed [14] known as Competitive linking algorithm (CLA). He aligned the word in one to one mapping by providing pre-existing knowledge (dictionaries).

Competitive linking algorithm (CLA): It is a greedy algorithm used for linking the corresponding text with each other. It calculates the highest score which is linked to the corresponding pair and then pops them from the search vector so that multiple linked words can be avoided. The scores of links will be calculated until a threshold value is reached. The advantage of CLA is easiness, not has too complicated steps.

### 3.2 Statistical based approach

It is the technique based on the sentence length, word position, word frequency etc. as it is not dependent upon the lexical resources. It consists of a hidden variable which defines the mapping from source to target sentence and from the learning data, a set of unknown parameters are derived by using EM algorithm. It includes probabilistic models like IBM, Hidden Markov model (HMM) and LEAF. These above models need a tool which are GIZA++, Natura Aligner Tool (NATool), etc.

Some of the methods used in this approach given below:

K-vec algorithm: It uses the concept of word frequency feature and word position to find the correct correspondences and this algorithm was proposed by [15]. It is basically for a different pair of language. This algorithm focused on similarity measures, as the text are divided into K-segments and for each word a k-length binary vector is created for indicating the presence and absence of a word. If the vector has "1" value for a word, that means it has a similar word in the different language, otherwise "0".

DK-vec Algorithm: It is the derived from K-vec algorithm, the concept behind the K-vec algorithm is that if the words are having their translation then the words will be in the same segment. The main disadvantage of K-vec algorithm is that it doesn't consider priori knowledge. To overcome this disadvantage, DK-vec algorithm (Dynamic K-vec algorithm) was proposed by [3]. A recency information is provided with parameters given in K-vec algorithm and Dynamic time wrapping is calculated, which provide the accuracy of pattern matching.

The approaches used for English-Hindi dialects are given below:

| AUTHOR | APPROACHES & TOOL Used | Precision and Recall | AER% and F-Score% | LIMITATIONS |
|---|---|---|---|---|
| Niraj Aswani et al. (2005) | A hybrid approach which aligns at sentence and word level.<br>It was done manually. | 77% and 67.79% | - | System doesn't solve the problem of subject-object-verb problem. |
| Niladri Chatterjee et al. (2006) | recency-vector based approach using the DK-vec Algorithm and Somers' Algorithm<br>It was done manually. | - | - | works on the small corpus. |
| Venkataramani et al. (2010) | Corpus augmented method with GIZA++ tool and NATools | - | 22.45%( reduces the Alignment Error Rate (AER) approximately to 5.96%.) | Failed for longer sentences. |
| Jyoti Srivastava et al. (2012) | IBM model 1 with POS tagging which was done manually | - | 50.14% and 49.86% | doesn't work for the large corpus |
| Jyoti Srivastava et al. (2014) | IBM1 model + Clause Identification +POS Tagger which was calculated manually | 50% and 55.46% | 47.41% and 52.59% | doesn't work for the large corpus, only solved the problem of longer sentences. |
| Jyoti Srivastava et al. (2015) | POS tagging with GIZA++ tool | 44.61% and 52.03% | 51.96% and 48.04% | Can't map one to many words. |

## IV CHALLENGES

**4.1 A Complex structure of Hindi language**: In English there are at most of the 7-8-word forms of a noun but in Hindi, it can be more than 40. Due to which one word of English can have many forms in the Hindi language according to the meaning of the sentence.

**4.2 Limited resources:** The word alignment requires the well-defined dictionary**.** The dictionary helps in identifying the correct pair of the word while aligning. Some resources are available, but it should be improved.

**4.3 Syntax (SVO-SOV) problem:** The English language follows the subject verb object pattern whereas Hindi follows a subject object verb pattern. This makes the task complex to align one language with another.

**4.4 Word alignment of Hybrid language:** People those who are uncomfortable with the English language use hybrid language such as Hinglish, the combination of Hindi and English. This is a tedious task to create a new dictionary for Hinglish language.

**4.5 Alignment of Articles:** The Hindi language doesn't consist of articles, thus task complexity increased. How to align Article of English language with the Hindi language.

## V CONCLUSION

In this paper, a survey is done in the field of English- Hindi word alignment. The proposed work identifies the approaches and challenges given by different researchers in the context of word alignment in the different corpus. Some of them used Lexicon based approach while other have used Statistical based approach. This paper also enlists the challenges to be handled with the help of algorithms used in the word alignment. But, the problem that occurs in English- Hindi corpus can't be resolved with the limited resource available for the Hindi language. Working in the same direction our future work will focus on the solution of those challenges so that the accuracy and performance get improved. Furthermore, we will investigate the long sentences that tarnish the performance in the English-Hindi parallel corpus.

## REFERENCES

[1] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, *19*(2), 263-311.

[2] The Wikipedia [Online]. Available: htttp://www.wikipedia.org/wiki/Bitext_word_alignment.

[3] Chatterjee, N., & Agrawal, S. (2006, July). Word alignment in English-Hindi parallel corpus using recency-vector approach: Some studies. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 649-656). Association for Computational Linguistics.

[4] Venkataramani, E., & Gupta, D. (2010, December). English-Hindi automatic word alignment with scarce resources. In Asian Language Processing (IALP), 2010 International Conference on (pp. 253-256). IEEE.

[5] Srivastava, J., & Sanyal, S. (2015, October). POS-based word alignment for the small corpus.In Asian Language Processing (IALP), 2015 International Conference on (pp. 37-40). IEEE.

[6] Gale, W. A., & Church, K. W. (1991, February). Identifying Word Correspondences in Parallel Texts. In HLT (Vol. 91, pp. 152-157).

[7] Cherry, C., & Lin, D. (2003, July). A probability model to improve word alignment. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume1 (pp. 88-95). Association for Computational Linguistics.

[8] Moore, R. C. (2004, July). Improving IBM word-alignment model 1. In Proceedings of the 42nd annual meeting on association for computational linguistics (p. 518). Association for Computational Linguistics.

[9] Aswani, N., & Gaizauskas, R. (2005, June). A hybrid approach to align sentences and words in English-Hindi parallel corpora. In Proceedings of the ACL Workshop on Building and Using Parallel Texts (pp. 57-64). Association for Computational Linguistics.

[10] Gao, Q., Vogel, S. (2008, June). Parallel implementations of word alignment tool. In Software Engineering, Testing, and Quality Assurance for Natural Language Processing (pp. 49-57). Association for Computational Linguistics.

[11] Srivastava J., Sanyal S. (2012) Segmenting Long Sentence Pairs to Improve Word Alignment in English-Hindi Parallel Corpora. In: Isahara H., Kanzaki K. (eds) Advances in Natural Language Processing. Lecture Notes in Computer Science, vol 7614. Springer, Berlin, Heidelberg.

[12] Pal, S., & Naskar, S. K. (2016). Hybrid word alignment. In Hybrid Approaches to Machine Translation (pp. 57-75). Springer International Publishing.

[13] Chen, S. F. (1993, June). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*(pp. 9-16). Association for Computational Linguistics.

[14] Melamed, I. D. (1997, July). A word-to-word model of translational equivalence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 490-497). Association for Computational Linguistics.

[15] Fung, P., & Church, K. W. (1994, August). K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics-Volume 2* (pp. 1096-1102). Association for Computational Linguistics.