A Hybrid Approach for Word Alignment in English-Hindi Parallel Corpora with Scarce Resources

Jyoti Srivastava
Information Technology
Indian Institute of Information Technology, Allahabad
Allahabad, India
Email: jyoti11s@gmail.com

Sudip Sanyal
Information Technology
Indian Institute of Information Technology, Allahabad
Allahabad, India
Email: ssanyal@iiita.ac.in

***Abstract*—This paper presents an approach which improves the performance of the word alignment with scarce resources for English-Hindi language pair. We obtain an improvement in the performance of IBM Model 1-2 algorithm by applying part of speech (POS) tag prior to the computation of word alignment probability. This paper demonstrates the increase of precision, recall and F-measure by approximately 15%, 11%, 14% respectively and reduction in Alignment Error Rate (AER) by approximately 14% with IBM Model 1. Similarly it shows an increase of precision, recall and F-measure by approximately 6%, 6% and 6% respectively and reduction in Alignment Error Rate (AER) by approximately 6% with IBM Model 2. Experiments of this paper are based on TDIL corpus.**

***Keywords*-Word alignment; Statistical Machine Translation; POS tagger; Scarce resources;**

I. Introduction

Word alignment is the task of identifying the correct translation relationships between the words of a parallel corpus [1] [2]. Many word alignment techniques have been developed so far in Natural Language Processing (NLP). But, word alignment techniques between English and Hindi did not have much progress due to the complex structure of the participating languages and scarcity of Hindi-language resources. This paper presents a simple method to improve the word-alignment accuracy of IBM Models in which these limitations have been overcome.

The basic hypothesis of this work is that if we apply a part of speech (POS) tagging prior to the calculation of the various probabilities, then the performance of word alignment will improve.

Word alignment models are a crucial component in statistical machine translation systems. IBM Model 1 is a word alignment model which is widely used for working with parallel bilingual corpora [1]. It was originally developed for providing reasonable parameter estimates to initialize more complex word-alignment models like IBM Models 2-5 and HMM. All the IBM models are relevant because EM training starts with the simplest IBM Model 1 for a few iterations and then proceeds through iterations of the more complex models all the way to IBM Model 5. So we can expect that if IBM Model 1 improves then higher IBM Models will also improve. One well known method of improving the accuracy is to increase the size of the parallel corpus. However, this is an expensive process. This paper is an effort to deal with word alignment when resources are scarce. We will work primarily with IBM Models 1 and 2.

II. Related Work

Word alignment is used in many applications. It is an essential step of statistical machine translation [1] [2]. Word aligned corpus is useful in automatic extraction of bilingual lexicon and terminology [3]. It can also help to transfer language tools developed for one language to other languages. Many NLP applications are enhanced and can improve their performance by using word alignment of better-quality [4].

Many word alignment techniques are proposed in the literature. Probabilistic generative approaches like IBM Model 1-5, HMM and LEAF are based on hidden alignment variable and they finally optimize word maps using EM algorithm [1] [5] [6]. Most practitioners still use IBM Models and HMM models for word alignment. The tool based on these methods is GIZA++. Word alignment is the first step of the statistical machine translation. HMM word alignment model and GIZA++ which is an implementation of the IBM Model 1-5 are the most widely-used word alignment system [1].

In particular, for English-Hindi word alignment, some work has been reported. Chatterjee and Agrawal have conducted experiments on manually lemmatized parallel corpus using recency vector based approach [7]. Aswani and Gaizauskas proposed hybrid approach based on local word grouping, cognates, nearest aligned neighbor, dictionary lookup, transliteration similarity and finally language dependent grammar rules for the alignment of English-Hindi bilingual corpus [8] [9]. Several recent works incorporate syntactic features into alignment to improve the performance of word alignment [10]. A generic discriminative re-ranking approach for English-Hindi word alignment is proposed which use structural features of both languages [11]. Venkataramani and Gupta provide a corpus-augmented method of word alignment for English-Hindi with scarce

978-0-7695-4886-9/12 $26.00 © 2012 IEEE
DOI 10.1109/IALP.2012.13

resources [12]; they used two existing word alignment tools: GIZA++ and NATools. Dhekane suggested the use POS tagger to create a bilingual dictionary [13].

III. Proposed Approach: IBM Model with POS Tagger

The conventional IBM Model 1 computes the alignment probability for each source word of the source sentence with each target word of the target sentence. The approach proposed in this paper computes the alignment probability of a source word of a source sentence with only those target words of the target sentence which have the same POS tag as the source word. Thus it reduces the computation time and improves the performance.

POS tags also solve the problem of word sense disambiguates at a small scale. For example: book will have different translations depending on whether it is used as noun or verb. In pure statistical technique like IBM Models, book will be translated to book as verb or as noun but not both. On the other hand, if the sentence containing the word book is POS tagged, then book will be tagged as verb or noun, depending on the sentence. Now, while looking for translations of the word book, the translation system will search for their corresponding translation with tag information using statistical techniques and it will find the right translation. Proposed approach is given in Algorithm 1 where format of the Sentence is T/w_1, T/w_2, ..., T/w_n where w_i is i_{th} word and T is POS tag of i_{th} word.

An unexpected problem is observed when analyzing the corpus after applying POS tag. In English-Hindi translation, a verb in English sentence often becomes a combination of noun and verb or adjective and verb in Hindi sentence.

For example:
Caring/VB - देखभाल/NN करना/VB
Promoting/VB - विकसित/JJ करना/VB
Simplifying/VB - सहज/JJ बनाना/VB

To solve this problem alignment probability of a verb in English sentence is computed with not only all the verbs of the Hindi sentence but also with the entire noun and adjective that immediately followed by the verb.

IV. Evaluation Metrics

This paper reports the performance of proposed approach in terms of two different measures: F-measure and alignment error rate (AER). These measures were also frequently used in the previous word alignment literature. AER is a measure of quality of word alignment and is defined by Och and Ney [4]. Alignment A is the set of alignments produced by the alignment model under testing. With a gold standard alignment G, each such alignment set consisting of two sets A_S, A_P and G_S, G_P corresponding to Sure (S) and Probable (P) alignments, these performance statistics are defined as

$$(Precision)P_T = \frac{|A_T \bigcap G_T|}{|A_T|} \tag{1}$$

Algorithm 1 Proposed Word Alignment Algorithm

procedure $WordAlignment(E, H)$
  for $i \leftarrow 1, n$ **do** ▷ n is the size of the corpus
    for each tag T of E and H **do**
      Create Seperate List $T_L(TAG, E_L, H_L)$ for each tag T ▷ TAG - POS tag, E_L - Source word list, H_L - Target word list
    end for
    for $j \leftarrow 1, length(E_i)$ **do**
      Extract T from T/e_j
      if T is already in T_L as TAG **then** ▷ If same tag exist in tag list
        Append e_j to corresponding list E_L
        Jump to Step 6.
      end if
      $TAG \leftarrow T$
      Append e_j to list E_L
      for $k \leftarrow 1, length(H_i)$ **do**
        Extract T from T/h_k
        if $T = TAG$ **then** ▷ If source and target word are of the same tag
          Append h_k to list H_L
          Remove h_k from the sentence H_i
          $k \leftarrow k - 1$
        end if
      end for
      if H_L is empty **then** ▷ If no target word have the same tag
        $H_L \leftarrow NULL$
      end if
    end for
    if H_i is not empty **then** ▷ if some tag of target word are not in source sentence
      for $k \leftarrow 1, length(H_i)$ **do**
        Extract T from T/h_k
        $TAG \leftarrow T$
        $E_L \leftarrow NULL$
        Append h_k to list H_L
      end for
    end if
  end for
  Now compute IBM Model 1 algorithm, but instead of computing for each sentence pair now compute for each tag pair T_L.
end procedure

$$(Recall)R_T = \frac{|A_T \bigcap G_T|}{|G_T|} \tag{2}$$

$$(F-measure)F_T = \frac{2P_T R_T}{P_T + R_T} \tag{3}$$

$$AER = 1 - \frac{|A_P \bigcap G_S| + |A_P \bigcap G_P|}{|A_P| + |G_S|} \tag{4}$$

Where T is the alignment type which can be set to either S or P.

V. Result and Discussion

We have tested our approach and the conventional IBM Models 1 and 2 on TDIL corpus. The algorithm of IBM Model 1 and IBM Model 2 used here is taken from the book of Statistical Machine Translation [14]. Fig. 1 and Fig. 2 shows the improvement in the performance of IBM Model 1 when the POS tag is used to compute the word alignment probability. These results demonstrate that the performance of the word alignment model improve when the size of the parallel corpus is increased. This is, of course, expected and is true for any statistical process i.e. the percentage error decreases as the sample size increases. The surprise observation is that even with a small corpus of 90 training sentences, but with the POS tags, the accuracy is comparable to that of the 270 training sentences.

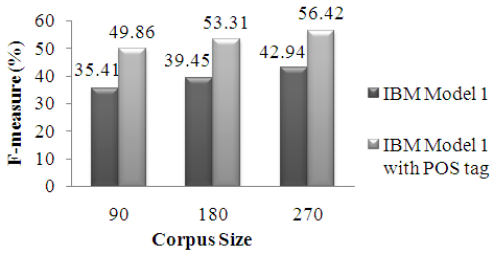

Figure 1. Comparison of F-measure of IBM Model 1 and proposed approach for different corpus size

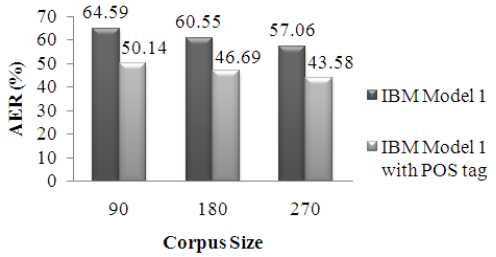

Figure 2. Comparison of AER of IBM Model 1 and proposed approach for different corpus size

Fig. 1 shows that the AER obtained using 270 training sentences (without POS tagging) is 57.06% while with the small corpus of 90 training sentences (with POS tagging) we get an AER of 50.14%. Similarly Fig. 2 shows that the F-measure obtained using 270 training sentences (without POS tagging) is 42.94% while with the small corpus of 90 training sentences (with POS tagging) we get F-measure of 49.86%. Thus we can see that the performance improves by 7% by using proposed approach even when we use one third

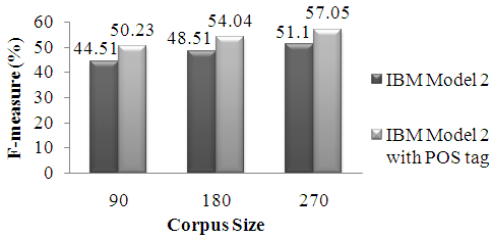

Figure 3. Comparison of F-measure of IBM Model 2 and proposed approach for different corpus size

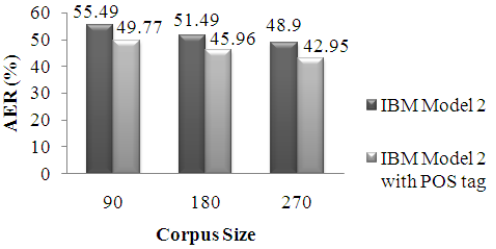

Figure 4. Comparison of AER of IBM Model 2 and proposed approach for different corpus size

(90 training sentences) of the complete training corpus (270 training sentences).

Fig. 3 and Fig. 4 shows the improvement in the performance of IBM Model 2 when POS tags are used to compute the word alignment probability. Fig. 3 shows that F-measure is increased by 6% on the same corpus when the proposed approach is used. Similarly Fig. 4 shows that AER is decreased by 6% on the same size of corpus when the proposed approach is applied.

Concluding from these results, only with extremely small bilingual training corpus and POS tagger for source and target language, word alignment model gives good results.

VI. Conclusion and Future Work

This paper presents a hybrid method of word alignment for English-Hindi with scarce resources. The proposed approach of POS tagging with word alignment model contributes significantly in the reduction of Alignment Error Rate (AER). All the conducted experiments prove that using POS tagger with any IBM Model performs better when compared to the use of any IBM Model 1-5 individually, for the task of word alignment. This experiment provides new avenues to extend this approach for other Indian languages. This paper demonstrated that it is possible to improve the performance of IBM Model 1 in terms of F-measure and AER by about 14%, simply by using POS tagger.

This paper focuses on developing suitable word alignment schemes in parallel texts where the size of the corpus is not too large. The paucity of the resources suggests that purely statistical techniques are not suitable for the task. A

deeper analysis of the error cases shows that F-measure can be increased and AER can be decreased further by providing the solution for the following problems:

- IBM Model 1-2 has problems of fertility, distortion and many words to one word translation. This problem can be solved by applying POS tagger on higher IBM Models which solve this type of problems.
- There are many long sentences in the corpus which degraded the performance. This problem can be solved by breaking the longer sentences into smaller one.
- There might be multiple Hindi equivalent word for the same English word which is the primary reason for poor performance in English-Hindi context. Due to which the frequencies of word occurrences differ significantly in the corpus and thereby jeopardize the calculations. This problem can happen due to three reasons:
  - Adjectives may have several declensions in Hindi but not in English. For example, for the English word black there are three different Hindi words "kaalaa (काला)", "kaalii (काली)", and "kale (काले)". The use of these words depends on the number and gender of the noun.
  - Nouns and pronouns can also have different declensions in Hindi. It depends on the case endings and/or the number and gender of the object. For example, the English word my has different translation in Hindi (e.g. "meraa (मेरा)", "merii (मेरी)", "mere (मेरे)"). Similar is true for nouns too. For example, the Hindi translation of the word hour is "ghantaa (घंटा)", while the plural form, hours can becomes "ghante (घंटे)" or "ghanto (घंटों)".
  - Morphology of verbs in a Hindi sentence depends on the gender, number and person of the subject of the sentence. In Hindi, there are 11 possible suffixes (e.g taa (ता), te (ते), tii (ती)) which might be attached to the root Verb of the sentence to render the morphological variations. For example, the word write in English is translated as ElKnA in Hindi, but depending on the gender, number and person of the subject of the sentence, write is translated as "likhataa (लिखता)", "likhatii (लिखती)", "likhate (लिखते)" and "likhegaa (लिखेगा)".

  Due to these problems, many English words are not correctly aligned.

ACKNOWLEDGMENT

We are grateful to IIIT Allahabad for providing the suitable infrastructure for research. This research has been funded by Tata Consultancy Services (TCS).We would like to thank Indian Language Technology Proliferation and Deployment Centre Team for providing Sample tourism parallel corpus. We are grateful to the organizers of ACL 2005 workshop to make the word alignment evaluation code as open source.

REFERENCES

[1] Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL, The mathematics of statistical machine translation: Parameter estimation. In Computational Linguistics, 1993, 19(2), pp. 263-311.

[2] Gale WA and Church KW., Identifying word correspondences in parallel texts. In Proceedings of the Fourth DARPA Workshop on Speech and Natural Language. Asilomar, 1991, pp. 152-157.

[3] Smadja F A., McKeown KR. and Hatzivassiloglou V, Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics, 1996, 22(1), pp. 1-38.

[4] Och FJ and Ney H, A systematic comparison of various statistical alignment models. In Computational Linguistics, 2003, 29(1), pp. 19-51.

[5] Vogel S, Ney H and Tillmann C, HMM-based word alignment in statistical translation. In Proceedings of Association of Computational Linguistic, 1996, Volume 2, pp. 836-841.

[6] Fraser A and Marcu D, Getting the structure right for word alignment: LEAF. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, 2007,pp. 51-60.

[7] Chatterjee N and Agrawal S., Word Alignment in English-Hindi Parallel Corpus Using Recency-Vector Approach: Some Studies. In Proceedings of the 21st International Conference on Computational Linguistics, Sydney, Australia, 2006, pp. 17-21.

[8] Aswani N and Gaizauskas R., 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, Ann Arbor, 2005, pp. 57-64.

[9] Aswani N and Gaizauskas R., Aligning words in English-Hindi parallel corpora. In Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, 2005, pp.115-118.

[10] Yanjun Ma, Patrik Lambert and Andy Way. Tuning Syntactically Enhanced Word Alignment for Statistical Machine Translation. In Proceedings of the EAMT 2009, the 13th Annual Meeting of the European Association for Machine Translation, Barcelona, Spain., 2009, pp. 250-257.

[11] Venkatapathy S and Joshi AK, Discriminative word alignment by learning the alignment structure and syntactic divergence between a language pair. In Proceedings of SSST, NAACL-HLT 2007 /AMTA Workshop on Syntax and Structure in Statistical Translation. Rochester, New York, 2007, pp. 49-56.

[12] Venkataramani E and Gupta D., English-Hindi Automatic Word Alignment with Scarce Resources. International Conference on Asian Language Processing, 2010, IEEE. pp. 253-256.

[13] Dhekane RM, Statistical Approach with Factored Translation Models for Indian Languages. Indian Institute of Technology, Kanpur, Advanced Natural Language Processing, 2009.

[14] Koehn P, Statistical Machine Translation. Book. Cambridge University Press, Published in the United States of America by Cambridge University Press, New York, 2010.