Maximum Likelihood Estimation for Bangla–Odia Word Alignment

Bishwa Ranjan Das, Dilip Singh, Prakash Chandra Bhoi, and Debahuti Mishra

Abstract In this paper, the mathematical function Maximum Likelihood Estimation (MLE) used to measure Bangla–Odia word alignment performance is used to provide better results and good accuracy. This MLE technique helps to find out the maximum likelihood probability value with the collaboration of the 'argmax function' that follows the mapping between two or more words of source and target language sentences. The lexical relationship among the words between two parallel sentences knows after calculating some mathematical values and those values indicate which word of the source language (SL) is aligned with which word of the target language (TL). Find MLE or MAP, the maximum a posterior parameter in the probability model, which depends on the unobserved latent model or hidden variables. Keeping all these issues in mind, it is described the nature of lexical problems that arise at the time of analyzing bilingual translated texts between Bangla (source language) and Odia (target language). The basic challenges lie in the identification of the single word units of the source text which are converted to multiword units in the target text and vice versa. The experimentally, there are thousands of parallel sentences are taken as training set and out of these sentences only hundreds of parallel sentence pairs are considered for test data. The accuracy of the proposed model is giving better performance as compared to other model and the accuracy which is more than expectation.

Keywords Expectation · Maximization · Probability · Alignment · Divergence · Odia · Bangla

B. R. Das (✉) · D. Singh · P. C. Bhoi · D. Mishra
Department of Computer Science and Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India

D. Mishra
e-mail: debahutimishra@soa.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Mallick et al. (eds.), *Electronic Systems and Intelligent Computing*, Lecture Notes in Electrical Engineering 860, https://doi.org/10.1007/978-981-16-9488-2_31

1 Introduction

Word alignment is the process of identifying or mapping the exact and corresponding word between two parallel corpora. It is one of the translation relationships of the words between two or more parallel sentences. Somehow, a word is translated by a single word or multiple words called word divergence. In the given parallel sentences, to find the corresponding relationship among words that may be one-to-one, one-to-many, and many-to-many of source and target sentences remains the main task of word alignment. Alignment of source language phrases with corresponding target language phrases or groups of words is the solution of phrase-based translation. If the phrases of the supply sentence are not able to discover their suitable translation withinside the goal language, clearly, they're assigned null. The movement of translated words in the source sentence to their appropriate position in the target sentence is also done in word alignment. In the case of bilingual machine translation, the word reordering may be a necessity and word alignment helps in achieving it. There are multiple factors for word alignment, i.e., Named entities, Transliteration similarities, Local word grouping, nearest aligned neighbors and dictionary lookup. The various challenges of achieving word alignment include ambiguity, word order, word sense, idioms, and pronoun resolution can be solved by mathematical operation and some conceptual concept of linguistics. In Word alignment, handling the 'Word divergence' or 'lexical divergence' problem is the main issue and challenging task here though it is not solved by many more algorithms till now it is only possible through a bilingual dictionary or called lexical database that is experimentally examined and tested only mathematically. Problems of word divergence or lexical divergence are normally addressed at the phrase level using bilingual dictionaries or lexical databases.

In the information of phrase alignment, the use of numerous the use of techniques inclusive of hybrid approach which plays nearby phrase grouping on Hindi sentences and makes use of different techniques which includes dictionary lookup, transliteration similarity, anticipated English phrases and nearest aligned neighbors. The probability values between small and large pair of sentences are discussed thoroughly [1]. The various issues, problems, and challenges are described very briefly here. Different types of approaches are also described thoroughly [2]. Most of the challenges are faced and solved very carefully using Expectation Maximization algorithm and using statistical technique, the whole concept is described very prominently with good accuracy. Most of the problems and issues are solved here [3]. In this paper, the various mapping techniques one-to-one, many-to-one are solved for Bangla–Odia lexical divergence problem [4]. In this paper, for estimating the parameters of those models given a hard and fast of pairs of sentences which can be translations of each other is defined through a sequence of five statistical models of the interpretation system and algorithms [5]. English–Hindi parallel words of the sentences are mapping using word dictionary [6]. Automatic word alignment has been done using different approaches like boundary detection approach, minimum distance function, and dictionary look up [7]. Compound word spitting is the most important part of machine translation which breaks the whole word into different meaning of the word.

Different approaches and their advantages and disadvantages are elaborated systematically as well as discussed, the challenges faced during translation of one language to another [8]. In this paper, a new probabilistic version is supplied for phrase alignment wherein phrase alignments are related to linguistically encouraged alignment sorts. A novel undertaking of joint prediction of phrase alignment and alignment sorts is being proposed and applied novel semi-supervised gaining knowledge of set of rules for this undertaking [9]. The algorithm illustrated with examples: pooling information from more than one noisy source and turning into an aggregate density [10]. A collection of five statistical version of translation method is defined and algorithms are given for estimating the parameters of those models additionally proven a fixed of pairs of sentences which are translation of each other and is described an idea of phrase-by-phrase alignment among such pairs of sentences [11]. This book provides a comprehensive and clear introduction to the most prominent techniques employed in the field of statistical machine translation [12]. Semantic relationship can be used to improve the word alignment, in addition to the lexical and syntactic feature that are typically used [13].

2 Estimation of Maximum Likelihood

MLE is a way that discover values for the parameters of a version. The parameter values are determined such that they maximize the probability that the method defined via way of means of the version produced the information have been really observe.

Maximum likelihood estimation is a technique of calculating the parameters of a possibility distribution technique via way of means of maximizing the possibility price the usage of argmax characteristic in order that the assumed statistical model, the discovered facts is maximum probable. The price withinside the parameter area that maximizes the chance characteristic is referred to as the most chance estimate. The good judgment of most chance is each intuitive and bendy to calculate the most price amongst all chance's values. Now it is mostly dominate the all maximization functions.

3 Word Alignment with Methodology

This paper presents to learn and implement conditional probability model between Bangla and Odia sentence, denoted as $P_\theta(B|O)$. If the alignment of the sentences is observed before, then only estimate the P(B|O) that means to find the MLE value by taking some sentence pairs as an example. The subscript θ represents set of parameters having a dataset D of n sentences pairs, D = {(B1, O1), (B2, O2), (B3, O3), ..., (Bn, On)}, where each subscript n indexes a different pair and it represents number of sentence pairs that means (B1, O1) is one pair, (B2, O2) is another pair and so on. The model is fully trained to predict the existence of the missing word alignment. These are many ways to define P(B|O). Suppose a Bangla sentence B is represented by an array of I, (B1, B2, B3, ...BI) and an Odia sentence O is represented by an array of J, (O1, O2, O3, ..., OJ). The Bangla–Odia word can be represented as an array of length I, is (a1, a2, a3, . . . ai) where a1, a2, a3, . . . ai one–one alignment variables are. An alignment variable ai takes a value in the range [0, J]. If ai = 0 means j value is also 0 because ai = j, that means Bi is not aligned to any word Odia word called null alignment. Consider the sentence pair Bangla–Odia as an example. But in this particular example there is no null value exist. It may be arise in other pair of sentences in the whole corpus.

Bangla sentence

রবিবার\N_NNP মায়চা\N_NNP গ্রামে\N_NN কৃষক\N_NNP সংঘর্ষ\N_NNP সমিতি\N_NNP পঞ্চায়েত\N_NN করে\V_VM_VNF ২৫\QT_QTC অক্টোবর\N_NNP থেকে\PSP নির্মাণ\V_VM_VNF কার্য\N_NN বন্ধ\N_NN করার\V_VM_VNF সিদ্ধান্ত\N_NN নিয়েছে\V_VM_VF ।\RD_PUNC

Transliteration. Rabibar Mayacha grame krushak sangharsh samite panchayate kobe 25 Octobar theke nirman karjya band korar sidhant niechhe.

Odia sentence

ରବିବାର\N_NNP ଦିନ\N_NN ମାୟାଚା\N_NNP ଗ୍ରାମରେ\N_NN କୃଷକ\N_NN ସଂଘର୍ଷ\N_NN ସମିତି\N_NN ପଞ୍ଚାୟତ\N_NN ବସାଇ\V_VM_VNF ୨୫\QT_QTC ଅକ୍ଟୋବରରୁ\N_NNP ନିର୍ମାଣ\N_NN କାର୍ଯ୍ୟ\N_NN ବନ୍ଦ\N_NNV କରାଇବାକୁ\V_VM_VINF ନିଷ୍ପତ୍ତି\N_NN ନେଇଛି\V_VM_VF ।\RD_PUNC

Transliteration. Rabibar dino mayacha gramare krushaka sangharsha samiti panchayata basai 25 Octobarru nirmana karjya band karaibaku nispatti neichhi.

The Bangla sentence is a length of 17 and the Odia sentence length is also 17. The Bangla sentence length indicates as I and so on the Odia sentence length indicates as J. The words of both the sentences are indexed like B1, B2, B3, . . . , BI and O1, O2, O3, ..., OJ. The value of an alignment array ‘a’ will be {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17}. These are all j values. It is being assumed that the probabilistic model automatically create the Odia sentence from Bangla using a normal method. First of all, the size of Odia sentence I is chosen as per the probability distribution P(I|J), i.e., P(17|17). Since the P(I|J) can be written mathematically as P(1, 2, 3, ...I|1, 2, 3, . . . J) i.e., P (length of source language followed by target language). Then each Bangla word position aligns to an Odia word (or null) according to the valid sentence alignment of the standard corpus (ILCI) is P(ai = j|J). Finally, each Bangla word Bi is translated according to the probability distribution function on the aligned Odia word, P(Bi|O_{a_i}). So, for this alignment, all probability values are multiplied likewise P (Rabibar dino|Rabibar), P(Mayacha|Mayacha), P(grama|grame), and so on. The joint probability value of the Bangla sentence and its alignment conditioned, both are calculated on the Odia sentence is simply the product of all these probabilities [15].

$$\mathrm{P}(B, a|O) = \mathrm{P(I|J)} \prod_{i=1}^{I} \mathrm{P}(a_i|\mathrm{J}).\mathrm{P}\big(B_i|O_{a_i}\big) \tag{1}$$

It is basically two values, P(I|J), for all pairs of sentence lengths I and J, and P(B|O) for all pairs of co-occurring Bangla and Odia words B and O.

$$\forall_{O,B} P(B|O) \in [0, 1] \tag{2}$$

$$\forall_{O} \sum_{B} P(B|O) = 1 \tag{3}$$

4 Use of Maximum Likelihood Estimation

To observe the alignment, just taking care of the P(B|O) and estimate the approximate value through maximum likelihood estimation (MLE). At first, the alignment of the sentence has been discussed properly before then start doing the word alignment between Bangla and Odia. But there is no such type of situation occurs in Bangla–Odia as it shows in French to English translation. For Example, most of the word of French is aligned with the English word many times but this type of situation also arises in Bangla–Odia sentence pairs. From the understanding point of view, an MLE function is introduced here to calculate the probability of the given parameters. Here is showing one example how P(B|O) is calculated,

$$\begin{aligned} \theta_1 &= \mathrm{P(krushaka|krushakder)} \\ &= \frac{\mathrm{count(krushaka,\ krushakder)}}{\mathrm{count(krushaka,\ krushakder), P(krushakamananka, krushakder),\ P(krushamanankara,\ krushader)}} \\ &= 1/(1+2+1) = {}^{1}\!/\!{}_{4} = 0.25 \end{aligned} \tag{4}$$

$$\begin{aligned} \theta_2 &= \mathrm{P(krushaka|krushakder)} \\ &= \frac{\mathrm{count(krushakamananka,\ krushakder)}}{\mathrm{count(krushakamananka\ krushakder), P(krushaka, krushakder),\ P(krushamanankara,\ krushader)}} \\ &= 2/(1+1+2) = 2/4 = {}^{1}\!/\!{}_{2} = 0.5 \end{aligned} \tag{5}$$

$$\begin{aligned} \theta_3 &= \mathrm{P(krushaka|krushakder)} \\ &= \frac{\mathrm{count(krushaka,\ krushakder)}}{\mathrm{count(krushaka,\ krushakder), P(krushakamananka, krushakder),\ P(krushamanankara,\ krushader)}} \\ &= 1/(1+2+1) = {}^{1}\!/\!{}_{4} = 0.25 \end{aligned} \tag{6}$$

From these three equations [14], the Bangla word “krushakder” is aligned with different Odia words many times with different probability values. The matter is which value should be chosen for consideration. Sometimes it’s depended on highest probability value as find out by MLE here. But three parameters θ_1, $\theta_{2,}$ and θ_3 have different values of different alignments. If the highest value is considered, i.e., 0.5(Eq. 4) for P(krushaka|Krushakder) not always satisfied for all cases, only satisfied for that particular semantic sense of the sentence. So, MLE is not always good at all for all cases to find the exact values.

$$\prod_{n=1}^{N} P_\theta\left(B^{(n)}, a^{(n)}|O^{(n)}\right) = \prod_{n=1}^{N} P(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} P(a_i^{(n)}|J^{(n)}.P(B_i^{(n)}|O_{a_i}^{(n)}) \qquad (7)$$

Here, N is number of sentences, the source length language Bangla is I, the target language Odia length is J, i is the alignment index, and a_i is the alignment.

Now data is observed, and the parameters are estimated, finally need a probability function to find the highest value as our data(value) is highly probable under this model.

$$\hat{\theta} = \underset{\theta}{argmax} \prod_{i=1}^{N} P_\theta(B^{(n)}, a^{(n)}\big|O^{(n)}) \qquad (8)$$

In Eq. (4), where $\hat{\theta}$ it searches the highest probability value of word alignment by argmax function for each and every word in a sentence. It is basically a searching problem from an infinite number of possible sentences in the case of machine translation. Only one sentence is selected from different possible sentences after translation in agreement with the corpus. For this case, though the search problem is trivial, because the solution for $\hat{\theta}$ when the data described by model is fully observed. An algorithm is developed to learn θ from our hypothetical aligned data actually initiates the strategy or model which is described here. The data is scanned and observing the alignments and counting them (means aligned data) for each Bangla–Odia word pair. To calculate the probabilities values (aligned word pair Bangla–Odia), all counts (means probability values) are normalized by the number of times that is observed the corresponding Bangla word participating in any alignment. This implies an algorithm which is described here.

Algorithm

Step 1. Initialize all counts to 0

Step 2. For each n value between 1 to N

Step 3. For each i value between 1 to I

Step 4. For each j value between 1 to J

Step 5. Compare $a_i = j$ upto n i.e. i value

```
Step 6. Count [($B_i$, $O_j$)] + +
Step 7. Count [$O_j$] + +
Step 8. For each ($B_i$, $O_j$) value in count do
Step 9. P(B|O) = Count(B,O)/Count(O)
```

This algorithm implements over all pairs of the word in each to collect count, a computation that's quadratic in sentence length. This is not strictly necessary: it could have just looped over the alignment variable to collect the counts, which is linear. However, thinking about the algorithm as one that examines all pairs of a word will be useful when it is moving to the case of unobserved alignments, which turns out to be an extension of this algorithm. Here, two formulae are used to calculate alignment probabilities after some iteration.

A Bangla sentence $B = b_1, b_2, b_3 \ldots . b_i$ and translated into an Odia sentence $O_= o_1, o_2, o_3 \ldots o_j$. Among all possible Odia sentences, one is looked for the highest probability P(B|O). Using Bayes' rule it may be written as follows:

$$P(O|B) = P(O)P(B|O)/P(B) \tag{9}$$

As the denominator is independent of O, finding the most probable translation e^* will lead to the noisy channel model for statistical machine translation.

$$e^* = \text{argmax } P(O|B) \tag{10}$$

$$= \text{argmax} P(O)(P(B|O) \tag{11}$$

where P(B|O) is the translation model and P(O) is referred to as the language model. In most of the cases, many-to-one and one-to-many world alignment is purely based on phrase-based translation, there is no other way to do translation when word divergence is seen in word alignment. A bilingual Bangla–Odia lexicon is developed as per the corpus based on the agriculture domain for mapping the words and translated very smoothly by one-to-one, one-to-many, and many-to-many.

5 Result and Discussion

In the bilingual dictionary based on the agriculture (Corpus collected from TDIL, Govt. of India) domain, a small handful of sentences (approximately five thousand), around fifty thousand words stored in a well-formatted and scientific manner for easy access with observed alignments. All observed alignments are trained and it produces a good estimate of θ as mentioned in Eq. (8). If we think as much as data, to get good estimates. It contains a one-to-one word, many-to-one, and many-to-one word correspondence. First of all, connections (as one-to-one mapping) are equally likely. After one iteration the model learns that the connection is made between most similar words from two parallel sentences by finding the probability value between 0 and 1. After another iteration, it becomes clear that a connection between previous similar words is more likely as the probability value of the current word. So, bigram and trigram are the best method to find the probability of the sentence along with the alignment among the words. All probability values are calculated using a bigram with MLE and argmax function in the form of a table/matrix. All probabilities values calculated by MLE with argmax function is not sufficient for the finding to exact alignment two parallel sentences Bangla–Odia. Taking more than thousands of parallel sentences, the accuracy is not so satisfactory by experimentally done. So further, it will be tested by Expectation Maximization (EM) algorithm to get the good accuracy value for proposed system. So here a better probability distribution is being progressed. This percentage value can be further enhanced by using EM algorithm in near future. But here the accuracy is calculated manually using the mathematically formula Precision, Recall, and F-Score measure to reach near the threshold value around more than 80%.

6 Conclusion and Future Work

When a translation is occurred from one language to another, first of all, if a parallel corpus is properly aligned in sentence level, then word by word is easily done by machine. Most of the problem is raised like one-to-many and many-to-one alignment which are solved by bilingual dictionary and phrase-based translation. A bilingual dictionary is made one-to-one, one-to-many, and many-to-one correspondence (Bangla–Odia) between two languages is created. Sometimes phrased-level translation is a more appropriate solution for word divergence occurrences. The MLE function is used for finding the most suitable word pair between two languages (Bangla–Odia) from where the highest probability value is taken. It also helps to translate word by word, phrase wise and finding the appropriate position of the word of the target language with good accuracy. Time complexity is one of the major factors when data is huge for word alignment as well as machine translation. So, care should be taken to obtain a better result; to optimize this, is a challenging task. Space complexity not be reduced as our data or corpus is huge, space should be increased for this as memory is concern, otherwise, any research work based on NLP or Data Science will be superficial.

References

1. Aswani N, Gaizauskas R (2005) Aligning words in English-Hindi parallel corpora. Assoc Comput Ling 19: 115–118
2. Das BR, Maringanti HB, Dash NS, Word alignment in bilingual text for Bangla to Odia Machine Translation. In: Presented in the international conference on languaging and translating: within and beyond on 21–23, Feb 2020, IIT Patna, India
3. Das BR, Maringanti HB, Dash NS, Challenges faced in machine learning-based Bangla-Odia word alignment for machine translation. In: Presented in the 42nd international conference of linguistic society of India (ICOLSI-42) on 10–12 Dec 2020, GLA University, Mathura, UP, India
4. Das BR, Maringanti HB, Dash NS, Bangla-Odia word alignment using EM algorithm for machine translation, published in the journal of Maharaja Sriram Chandra BhanjaDeo (erstwhile North Orissa) University, Baripada, India
5. Brown PF, et al. (1993) The mathematical of statistical machine translation: parameter estimation. Comput Ling 19(2):263–311
6. Dubey S, Diwan TD (2012) Supporting large English-Hindi parallel corpus using word alignment. Int J Comput Appl 49:16–19
7. Jindal K, et al. (2011) Automatic word aligning algorithm for Hindi-Punjabi parallel text. In: International conference on information systems for Indian languages, pp 180–184
8. Koehn P, Knight K, Empirical methods for compounding splitting. EACL '03 Association for Computational Linguistics, vol 1, pp 187–193, April 12–17, 2003
9. Mansouri AB, et al. (2017) Joint prediction of word alignment with alignment types. Trans Assoc Comput Ling 5:501–514
10. Minka T (1998) Expectation-maximization as lower bound maximization
11. Della Pietra VJ, Della Pietra SA, Brown PF, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Ling 19(2):263–311
12. Koehn P (2010) Statistical machine translation
13. Songyot T, Chiang D (2014) Improving word alignment using word similarity. In: Empirical methods in Natural Language Processing, pp 1840–1845
14. Bhattacharyya P (2015) Machine translation. CRC Press, Taylor & Francis Group
15. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.5497&rep=rep1&type=pdf