SCIENCE SET
OPEN ACCESS PUBLISHERS

Research Article

Novel Journal of Applied Sciences Research

Word Alignment in Statistical Machine Translation: Issues and Challenges

Bishwa Ranjan Das[1], & Rekhanjali Sahoo[2]

[1]*Department of Computer Science and Engineering, NM Institute of Engineering and Technology, Bhubaneswar*
[2]*Department of Computer Science and Engineering, GITA Autonomous College, Bhubaneswar*

[*]**Corresponding author:** Bishwa Ranjan Das, Department of Computer Science and Engineering, NM Institute of Engineering and Technology, Bhubaneswar.

Submitted: 28 November 2024 **Accepted:** 05 December 2024 **Published:** 13 December 2024

Citation: *Bishwa Ranjan Das, & Rekhanjali Sahoo. (2024). Word Alignment in Statistical Machine Translation: Issues and Challenges. Nov Joun of Appl Sci Res, 1(6), 01-03.*

Abstract

Statistical Machine Translation (SMT) in the form of text is an approach to machine translation that is derived from the concept of information theory. Translating one natural text (Bangla) to another (Odia) using the learning technique is a challenging task with an issue called word alignment. Word alignment decides which word of source language (Bangla) is mapped with which word of the target language (Odia) using probability distribution values. The probability values are found by the iteration process between the words in a parallel corpus. A bilingual dictionary and phrase-based translation are sometimes required. A total of 70% of the text is taken for training, and 30% is taken for testing in the agriculture-based domain, which is collected from TDIL (Technology Development for Indian Languages, Govt. of India) and treated as the bilingual corpus or parallel corpus. The accuracy is calculated from the test set using the confusion matrix along with precision, recall, and f-score. That accuracy value indicates the performance of the model and will be enhanced further in the future. The accuracy is measured in three steps, including word-wise, phrase-wise, sentence-wise, and paragraph-wise translation, which give 0.92%, 0.88%, 0.85% and 0.83% respectively.

Keywords: Odia, Bangla, Corpus, Probability, Bilingual

Introduction

Word alignment is an important issue in statistical machine translation which is useful for mapping between source language text (Bengali) and target language text (Odia). It is the process of locating corresponding word pairs in two languages. There are some issues, where either a word is not translated or word is translated by multiple words. If parallel sentences are given then, finding the correspondences that are one-to-one, one-to-many and many-to-many between the words of source and target sentences are the main tasks of word alignment. Alignment of source language phrases with corresponding target language phrases or words is complicated for word alignment. If the words of source sentences are unable to find their appropriate translation in target language, they are assigned null. The movement of translated words in source sentence to their appropriate position in target sentence is also done by word alignment. In 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 machine translation include ambiguity, word order, word sense, idioms, and pronoun resolution [1-3].

Word Alignment

Word alignment remains a fundamental problem for statistical machine translation. Word alignment is typically done after sentence alignment. All current approaches use word alignment at some point during training or in feature functions. An alignment a=a1, a2, a3…ai is a vector of alignment variable, where each aj can take any value in the set {1, 2, 3,….,J}. The alignment vector specifies the mapping for each Bangla word to a word in the Odia sentence. ai=j, specifies that Bi is aligned to Oj.

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

୨୫\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 version – Ravibar mayacha grame krushak sangharsh samiti panchayate kabe 25 october theke nirman karjya band karaa sidhanta 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 version – Rabibar dino mayacha gramare krushaka sangharsha samiti panchayata basai 25 octoberru nirmana karjya band karaibaku nispati neichhi.

Here each word of Bangla sentence is one to one corresponding with Odia sentence except first word as per the corpus based on Agriculture domain. The first word of Bangla i.e. রবিবার (Rabibar) aligned with two word of Odia i.e. ରବିବାର ଦିନ (Rabibar dino). So, this type of concept is known as one-to-many relationship. Another problem, many-to-one is occurring here i.e. অক্টোবর থেকে – ଅକ୍ଟୋବରରୁ), two words in Bangla become one word in Odia according to parallel corpus. This problem can be solved only by phrase level translation.

Methodology

An alignment a=a1, a2, a3 ….. ai is vector of alignment variable, where each ai can take any value in the set {1, 2, 3,…., j}. The alignment vector specifies the mapping for each Bangla word to one or more word(s) in the Odia sentence and vice versa. If ai=0, this means that the Odia word is not aligned to any Bangla word; called null assignment. While ai=j, specifies that Bj is aligned with Oj. In the above example, the alignment pairs are a1=1,2, a2=3, a3=4, a4=5, a5=6, a6=7, a7=8, a8=9, a9=10, a10-11=11, a12=12, a13=13, a14=14, a15=15, a16=16, a17=17. Here a1, a2, a3…. represent Bangla words aligned with Odia words with index value 1, 2, 3…, of a sentence. It is being observed that the alignments are one-to-many, many-to-one and one-to-one i.e. more than one Bangla word aligned to one or more Odia word for this example strictly. The value of an alignment a will be {1,2,3,4,5,6,7-8,9,10-11,12,13,14,15,16,17}. So, the probabilistic model generates the Odia sentence from Bangla using a simple procedure. First the length I is chosen according to a distribution P(I|J), in this case P(17|17). Then each Bangla word position aligns to a Odia word (or null) according to the valid sentence alignment of 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(Oj|B_(a_i)). So, for this alignment, all probability values are multiplied. The joint probability of Odia sentence and its alignment conditioned on the Bangla is simply the product of all these probabilities [3-6].

$$P(O, a|B) = P(I|J) \prod_{i=1}^{I} P(a_i|J). P\left(O_j|B_{a_i}\right) \quad (1)$$

It is simply two tables of numbers: P(I|J), for all pairs of sentence lengths I and J; and P(O|B) for all pairs of co-occurring Odia and Bangla words O and B. since these numbers represent probabilities, the set of valid assignments of numbers to these tables must follow basic rules of probability.

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

$$\forall_B \sum_O P(O|B) = 1 \quad (3)$$

Now data is observed, and the parameters are to be estimated, needing a probability function to find the highest value.

$$\hat{\theta} = \underset{\theta}{argmax} \prod_{i=1}^{N} P_\theta(O^{(n)}, a^{(n)}|B^{(n)}) \quad (4)$$

In the equation 4, it searches the highest probability values for each and every sentence. It is basically a searching problem from infinite number of possible sentences. 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 there is a closed-form solution for $\hat{\theta}$ when the data described by our model is fully observed.

A Bangla sentence B=b1, b2, b3….bi and translated into an Odia sentence O=o1, o2, o3…oj. Among all possible Odia sentences, one is looked for the highest probability P(O|B). Using Bayes' rule, it can be written as

$$P(O|B) = P(O)P(B|O)/P(B) \quad (5)$$

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

$$e^* = \text{argmax}P(O|B) \quad (6)$$

$$= \text{argmax}(P(B|O) * P(O) \quad (7)$$

where P(O) is called the language model or prior probability, and (P(B|O) is the translation model or Word Alignment or likelihood probability. 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 Agriculture domain for mapping the words and translated very smoothly by one-to-one correspondence.

Result Discussion

In the bilingual dictionary based on agriculture domain, around fifty thousand words are stored in a well formatted and scientific manner (in Unicode or electronic format) for easy access. It contains one-to-one, one-to-many and many-to-one word correspondence. First of all, connections (one-to-one mapping) are equally likely. After one iteration, the model learns that the connection is

made between most similar words in two parallel sentences by finding the probability value. After another iteration, it becomes clear that a connection between previous similar words is more likely as the probability value of current word. So, Bigram is the best method to find the probability of the alignment among the words. All probability values are calculated using bigram in the form of table/matrix. Expected count, revised expected count and revised alignment probabilities values are calculated among words in each parallel sentence. Revised alignment probabilities give more approximation value between the words in the parallel sentence. The accuracy values of nearly 85% are calculated, taking fifty thousand words in a bilingual dictionary. This percentage value can be further enhanced by using other mathematical algorithms because day by day new algorithms are developed and we have test using new algorithm to enhance the accuracy values [6-10].

Conclusion and Future work

When translation is done from one language to another, first of all a parallel corpus is properly aligned in sentence level then continues with each individual word. Most of the problems arise like one-to-many and many-to-one alignment, which are solved by bilingual dictionary and phrase level translation. A bilingual dictionary with one-to-one correspondence (Bangla-Odia) between two languages is created. Sometimes, phrase level translation will be an appropriate solution, especially, in case of divergence occurrences. In this paper, Expectation and Maximization algorithm is used for finding the most suitable word pair between two languages Bangla-Odia. It also helps to translate word by word, phrase wise and finding the appropriate position of the word of target language. Time complexity is the one of the major factors when data is huge for translation in case of machine translation. So, care should be taken to obtain better results; to optimize this, is a challenging task [11-15].

References

1. Aswani, N., & Gaizauskas, R. (2005). Aligning words in English-Hindi parallel corpora. Association for Computational Linguistics, 19, 115–118.
2. Das, B. R., Maringanti, H. B., & Dash, N. S. (2016, March 25–28). Approaches to machine translation: A short review. Presented at the 3rd National Language Conference (NLC 2016), Institute of Odia Studies and Research (IOSR), Bhubaneswar, India.
3. Das, B. R., Maringanti, H. B., & Dash, N. S. (2017, July 6–9). Developing a transliteration system from English to Odia by using a statistical method. Presented at the 4th National Language Conference (NLC 2017), National Institute of Science and Technology, Berhampur, Odisha, India.
4. Das, B. R., Maringanti, H. B., & Dash, N. S. (2017, December 8–11). English-Odia machine transliteration system using probabilistic approach. Presented at the 39th International Conference of Linguistic Society of India (ICOLSI-39), Indian Institute of Technology, Patna, India.
5. Das, B. R., Maringanti, H. B., & Dash, N. S. (2020, February 21–23). Word alignment in bilingual text for Bangla to Odia machine translation. Presented at the International Conference on Languaging and Translating: Within and Beyond, IIT Patna, India.
6. Das, B. R., Maringanti, H. B., & Dash, N. S. (2020, December 10–12). Challenges faced in machine learning-based Bangla-Odia word alignment for machine translation. Presented at the 42nd International Conference of Linguistic Society of India (ICOLSI-42), GLA University, Mathura, UP, India.
7. Das, B. R., Maringanti, H. B., & Dash, N. S. (n.d.). Bangla-Odia word alignment using EM algorithm for machine translation. Journal of Maharaja Sriram Chandra BhanjaDeo (erstwhile North Orissa) University, Baripada, India.
8. Brown, P., et al. (1990). A statistical approach to machine translation. Computational Linguistics, 16, June.
9. Brown, P., et al. (1993). The mathematics of statistical machine translation: Parameters estimation. Computational Linguistics, 19(2), 261–311.
10. Dash, N. S. (2009). Translation corpora and machine-aided translation. Translation Today, 6(1 & 2), 134–153.
11. Dubey, S., & DharDiwan, T. (2012). Supporting large English-Hindi parallel corpus using word alignment. International Journal of Computer Applications, 49, 16–19.
12. Jindal, K., et al. (2011). Automatic word aligning algorithm for Hindi-Punjabi parallel text. Presented at the International Conference on Information Systems for Indian Languages, 180–184.
13. Kaur, H., & Laxmi, V. (2013). A survey of machine translation approaches. International Journal of Science, Engineering and Technology Research (IJSETR), 2(3), March.
14. Koehn, P., & Knight, K. (2003). Empirical methods for compounding splitting. EACL '03 Association for Computational Linguistics, 1, 187–193, April 12–17.
15. Mansouri, A. B., et al. (2017). Joint prediction of word alignment with alignment types. Transactions of the Association for Computational Linguistics, 5, 501–514.