REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 10 | ISSUE - 5 | FEBRUARY - 2021

भारतीय भाषा के लिए पदबंध आधारित सांख्यिकीय मशीन अनुवाद प्रणाली की तकनीकों और उपकरणों की समीक्षा

धीरेन्द्र यादव[1] डॉ. पीयूष प्रताप सिंह[2]
[1]शोधार्थी सूचना एवं भाषा अभियांत्रिकी केंद्र, म. गा. अं. हिं. वि., वर्धा, महाराष्ट्र, भारत.
[2]एसोसिएट प्रोफेसर, स्कूल ऑफ कंप्यूटर एंड सिस्टम साइंस, जवाहरलाल नेहरू विश्वविद्यालय, नई दिल्ली

सारांश (Abstract)

इस शोध–पत्र के माध्यम से भारतीय भाषाओं के लिए पदबंध आधारित सांख्यिकीय मशीन अनुवाद प्रणाली की तकनीकों एवं उपकरणों का एक सर्वेक्षण प्रस्तुत किया गया है। मशीन अनुवाद कंप्यूटेशनल भाषाविज्ञान के सबसे महत्वपूर्ण अनुप्रयोगों में से एक है जो एक भाषा से दूसरी भाषा में पाठ का अनुवाद करने के लिए कंप्यूटर सॉफ्टवेयर या वेब का उपयोग करता है। भारत एक बहुभाषी देश है। यहाँ कई राज्यों की अपनी मूल भाषा और लिपियाँ हैं। इसके लिए भारतीय भाषाओं के लिए स्वचालित मशीन अनुवाद प्रणाली की आवश्यकता है ताकि लोगों के बीच अपनी स्थानीय भाषा में जानकारी का आदान–प्रदान हो सके। इस शोध–पत्र में विभिन्न भारतीय भाषाओं में अनुवाद करने वाली पदबंध आधारित सांख्यिकीय मशीन अनुवाद प्रणालियों की समीक्षा की गई है ताकि भविष्य के अनुसंधान में और किसी विशेष उद्देश्य के लिए इन तकनीकों का उपयोग किया जा सके।

मूल शब्द (Keywords):– कंप्यूटेशनल भाषाविज्ञान, पदबंध आधारित मशीनी अनुवाद।

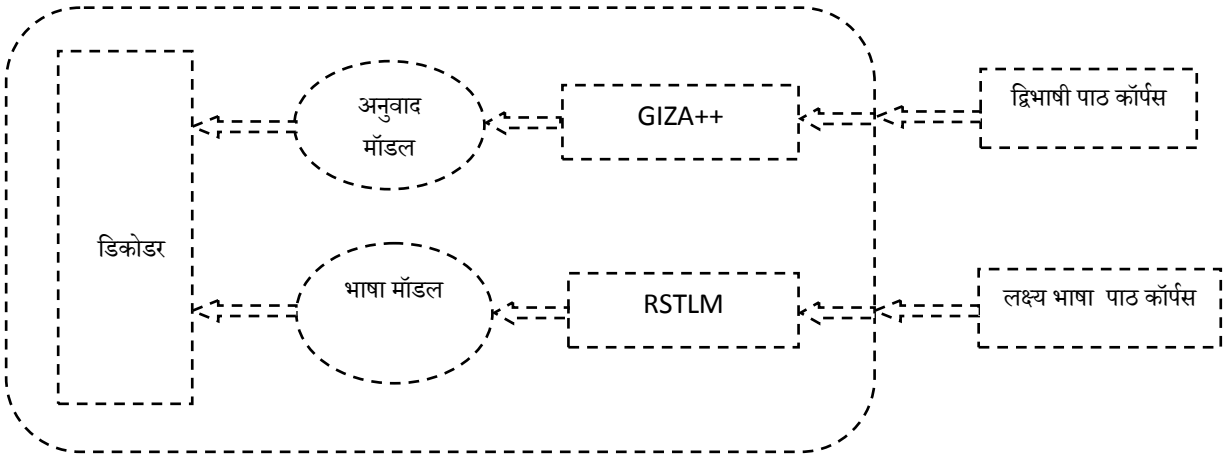
I. परिचय (Introduction):

मशीनी अनुवाद एक भाषा से दूसरी भाषा में पाठ के स्वचालित अनुवाद को संदर्भित करता है। मशीन अनुवाद प्रणाली कृत्रिम बुद्धि के प्राकृतिक भाषा संसाधन (NLP) का अनुप्रयोग है। मशीनी अनुवाद (Machine Translation) लोगों के बीच भाषा के अवरोध को तोड़ने के लिए महत्वपूर्ण होती है। भारत में कई शोधकर्ताओं, संस्थानों और संगठनों ने भारतीय भाषाओं के लिए मशीन अनुवाद प्रणाली पर काम करना शुरू कर दिया है तथा संतोषजनक परिणाम भी प्राप्त हुए हैं। भारत में मशीन अनुवाद को लेकर किए जा रहे अनुसंधान अपेक्षाकृत प्रारंभिक अवस्था में हैं। मशीन अनुवाद पर भारत के IIT कानपुर, IIT मुंबई, IIIT हैदराबाद, हैदराबाद केंद्रीय विश्वविद्यालय और C–DAC पुणे जैसे संस्थानों मे कार्य चल रहा है जिन्होनें इन प्रणालियों को विकसित करने में प्रमुख भूमिका निभाई है। भारत में पिछले दो दशकों से मशीनी अनुवाद के क्षेत्र में काफी शोध हो रहे हैं। कंप्यूटेशनल भाषाविज्ञान और कृत्रिम बुद्धिमत्ता में प्रगति के कारण 1990 के दशक के दौरान कुशल अनुवाद प्रणालियाँ बाजार में हैं। चूँकि इसमें और भी भाषाएँ हैं और विभिन्न भाषा बोलने वालों के बीच संचार सेतु उपलब्ध कराने के लिए अभी और काम चल रहा है। आज का दौर सूचना प्रौद्योगिकी का है

और इस दौर में प्रत्येक व्यक्ति अपनी निजी भाषा में इसका उपयोग करना चाहता है। इसके लिए प्राकृतिक भाषा संसाधन तकनोकी द्वारा कई सॉफ्टवेयर उपकरण विभिन्न क्षेत्रीय भाषाओं हेतु विकसित हो रहे हैं। विश्व में तेजी से हो रहे संगणकीय भाषा के विकास ने प्रत्येक भाषा की विशेषता को बारीकी से समझने पर ज़ोर दिया है। आज भारत के प्रत्येक गांव को एक साथ जोड़ने के लिए वैश्विक ग्राम की संकल्पना को साकार करने में सरकार भी प्रयत्नशील है। भारत वह राष्ट्र है जहाँ बोली जाने वाली भाषा में विविधता के साथ संस्कृति में बड़ी विविधता देखी जाती है। भारत जैसे बड़े बहुभाषी समाज में, एक भाषा से दूसरी भाषा में दस्तावेजों के अनुवाद की बहुत मांग है। अधिकांश राज्य सरकारें संबंधित क्षेत्रीय भाषाओं में काम करती हैं जबकि केंद्र सरकार के आधिकारिक दस्तावेज और रिपोर्ट द्विभाषी होते हैं अत: एक उचित संचार स्थापित करने के लिए संबंधित क्षेत्रीय भाषाओं में इन दस्तावेजों और रिपोर्टों का अनुवाद करने की आवश्यकता है। आजकल मशीन अनुवाद प्रणाली भारत में शोधकर्ताओं के लिए अध्ययन का एक उभरता हुआ क्षेत्र है। सीमित इलेक्ट्रॉनिक संसाधनों और उपकरणों के साथ किसी भी दो प्राकृतिक भाषाओं के लिए एक पूर्ण द्विभाषी मशीन अनुवाद प्रणाली का विकास एक चुनौतीपूर्ण कार्य है। पदबंध आधारित सांख्यिकीय अभिगमों का उपयोग करके विभिन्न भाषाओं के लिए मशीन अनुवाद प्रणाली विकसित करने के लिए दुनिया भर में कई प्रयास किए जा रहे हैं।

II. सांख्यिकीय मशीन अनुवाद (Statistical Machine Translation)

सांख्यिकीय मशीन अनुवाद एकभाषी और द्विभाषी प्रशिक्षण डेटा के विश्लेषण से उत्पन्न सांख्यिकीय अनुवाद मॉडल का उपयोग करता है। अनिवार्य रूप से यह दृष्टिकोण एक स्रोत भाषा को दूसरी भाषा के पाठ में अनुवाद करने के लिए परिष्कृत डेटा मॉडल बनाने के लिए कंप्यूटिंग शक्ति का उपयोग करता है। अनुवाद को प्रशिक्षण डेटा से एल्गोरिदम का उपयोग करके सबसे अधिक बार होने वाले शब्दों या वाक्यांशों का चयन करने के लिए चुना जाता है।



चित्र:1 सांख्यिकीय मशीन अनुवाद

- **भाषा मॉडल (Language Model)-** भाषा मॉडल का उद्देश्य एक वाक्य की गणना करके उसकी संभावना को विकसित करना है। भाषा मॉडल को एक शब्द की प्रायिकता की गणना के रूप में समझा जा सकता है, जो उस शब्द के सभी शब्दों को एक वाक्य में रखता है। भाषा मॉडल N–gram का उपयोग करके वाक्य की संभावना को बताता है। भाषा मॉडल एक वाक्य P (S) को तोड़ता है।
  $P(s)=P(w1,w2,w3,.....,wn)$
  $=P(w1)P(w2|w1)P(w3,|w1w2)P(w4|w1w2w3)...P(wn|w1w2...wn-1)$
- **अनुवाद मॉडल (Translation Model)** – अनुवाद मॉडल संभाव्यता P (T | S) की गणना करने में मदद करता है। अनुवाद मॉडल लक्ष्य भाषा और स्रोत भाषा के समानांतर कॉर्पस का उपयोग करके प्रशिक्षित किया जाता है। यह वाक्य को छोटी इकाइयों जैसे– शब्दों या वाक्यांशों में तोड़ता है। स्रोत वाक्य का लक्ष्य

अनुवाद शब्द से स्रोत शब्द से उत्पन्न होने के रूप म माना जाता है। उदाहरण के लिए इनपुट वाक्य S और उसके अनुवाद T को दर्शाने के लिए नोटेशन (T / S) का उपयोग करता है।

- **डिकोडिंग (Decoder)** – डिकोडिंग अनुवाद मॉडल और भाषा मॉडल का उपयोग करके स्रोत वाक्य के लिए एक लक्ष्य अनुवाद वाक्य को खोजने की एक प्रक्रिया है। जो अनुवाद और भाषा मॉडल की संभावना को अधिकतम करती है। इसमें उन शब्दों और वाक्यांशों को चुना जाता है। जिनमें अनुवादित अनुवाद होने की अधिकतम संभावना होती है। खोज वाक्य T को दर्शाता है जो P (S|T) को अधिकतम करता है।

$$\Pr(S,T) = \underset{\text{Decoding Algorithm}}{\text{argmax}} \underset{\text{LM}}{P(T)} \underset{\text{TM}}{P(S|T)}$$

यह संभावना को अधिकतम करने के लिए argmax () फ़ंक्शन का उपयोग करता है।

- **वाक्यांश आधारित मॉडल (Phrase based model)-** वाक्यांश आधारित मॉडल सबसे व्यापक रूप से उपयोग किए जाने वाला मशीनी अनुवाद दृष्टिकोण है, जिसमें शब्दों के छोटे अनुक्रम का एक बार में अनुवाद किया जाता है। वाक्यांश आधारित मशीन अनुवाद का मुख्य उद्देश्य यह है कि किसी भाषा के एक शब्द दूसरी भाषा में दो शब्दों के अनुरूप हो सकता है ताकि शब्द आधारित मॉडल ऐसे मामलों में विफल हो सकें। इसके अलावा वाक्यांश आधारित मॉडल अनुवाद तालिका में अधिक बेहतर स्थानीय पुन: व्यवस्थित और अधिक संदर्भ प्रदान करते हैं। MOSES और Phrasal दो सबसे व्यापक रूप से उपयोग किए जाने वाले फ्रेमवर्क हैं जो वाक्यांश आधारित मशीन अनुवाद के लिए समर्थन प्रदान करते हैं।

III. **पदबंध आधारित मशीन अनुवाद (Phrase&based Machine Translation):** पदबंध–आधारित सांख्यिकीय मशीन अनुवाद (PBSMT) मशीन अनुवाद के अन्य सभी तरीकों में सबसे लोकप्रिय दृष्टिकोण है। इस वाक्यांश आधारित SMT में एक भाषा मॉडल, एक अनुवाद मॉडल और एक विरूपण मॉडल होते हैं। गणितीय रूप से इसे इस प्रकार व्यक्त किया जा सकता है:

$$\text{ebest} = \text{argmax}_e P(e|f) = \text{argmax}_e [P(f|e) P_{LM}(e)]$$

यह एक भाषा मॉडल p (e) और एक अलग अनुवाद मॉडल p (f |e) के लिए अनुमति देता है।

IV. साहित्य सर्वेक्षण

1) Md, Zahurul Islam et al. (2010) ने एक वाक्यांश–आधारित सांख्यिकीय मशीन अनुवाद (SMT) प्रणाली का वर्णन करते हैं जो बांग्ला में अंग्रेजी वाक्यों का अनुवाद करता है। एक ट्रांसफ़ॉर्मेशन मॉड्यूल बाहर के शब्दावली (out–of–vocabulary) को संभालने के लिए जोड़ा जाता है। इस प्रणाली का कुल BLEU स्कोर 11.7 है। इस अनुवाद प्रणाली की गुणवत्ता में सुधार करने के लिए बेसलाइन अनुवाद प्रणाली (लिप्यंतरण और प्रीपोज़िशन हैंडलिंग) में दो अतिरिक्त मॉड्यूल शामिल किए गए है। जिसमें यह भी दिखाया गया है कि वाक्यांश–आधारित सांख्यिकीय मशीन अनुवाद मॉडल के साथ एक स्वचालित लिप्यंतरण प्रणाली का भी निर्माण किया जा सकता है। अनुवाद शुद्धता में सुधार करने के लिए प्रीपोज़ हैंडलिंग मॉड्यूल काफी प्रभावी है। कुल मिलाकर छोटे वाक्यों (23.30 BLEU और 0.63 TER) के लिए उचित अंक प्राप्त किए गए हैं।

2) Singh et al. (2012) ने इस शोध–पत्र में अंग्रेजी से हिंदी भाषा में पदबंध आधारित सांख्यिकीय मशीन अनुवाद प्रणाली का वर्णन किया है। जो डेटा संग्रह का पूर्व–प्रसंस्करण, मॉडलिंग, प्रशिक्षण और खोज से लेकर सभी पहलुओं में एक सांख्यिकीय मशीन अनुवाद प्रणाली के विकास का विस्तार से वर्णन करता है

तथा विकासवादी रैपिड प्रोटोटाइपिंग प्रतिमान का उपयोग करके विभिन्न सफल सांख्यिकीय मशीन अनुवाद विकसित किए गए हैं।

3) Joshi et al. (2013) के द्वारा 'Making headlines in Hindi' नामक एक अनुवाद इंजन प्रस्तुत किया गया है। जिसका उद्देश्य अंग्रेजी समाचारों के शीर्षकों की शैली को संरक्षित रखते हुए उनका हिंदी में अनुवाद करना है। इस इंजन में पदबंध–आधारित एवं कारक–आधारित दो मॉडल को सम्मिलित किया गया है। पदबंध–आधारित मॉडल एक डोमेन भाषा मॉडल और द्विभाषी शब्दकोश का उपयोग करता है। कारक–आधारित मॉडल POS, लेम्मा, काल और संख्या जैसे कारकों का उपयोग करता है। एक लोकप्रिय अंग्रेजी दैनिक द हिंदू 11 की वेबसाइट से डाउनलोड किए गए 787 सुर्खियों के परीक्षण सेट का उपयोग करके इंजन का मूल्यांकन किया गया और मूल वक्ताओं द्वारा हिंदी में मैन्युअल रूप से अनुवाद किया गया। वाक्यांश–आधारित मशीन अनुवाद के लिए BLEU स्कोर 13.40 और कारक–आधारित मशीन अनुवाद के लिए 5.73 प्राप्त किया गया।

4) Baruah et al. (2014) ने अपने इस शोध–पत्र में असमिया और अंग्रेजी भाषा के समानांतर कॉर्पस का उपयोग कर एक पदबंध आधारित सांख्यिकीय मशीनी अनुवाद प्रणाली प्रस्तावित की है जो असमिया से अंग्रेजी एवं अंग्रेजी से असमिया में अनुवाद करती है। इस प्रणाली के निर्माण हेतु पदबंध आधारित मोजेज़ टूलकिट का प्रयोग किया गया है। भाषा मॉडल विकसित करने और शब्दों को संरेखित करने के लिए इसमें क्रमशः दो अन्य उपकरण IRSTLM एवं GIZA का उपयोग किया गया है। प्रणाली को प्रशिक्षण करने के लिए लगभग 2500 वाक्यों का उपयोग किया गया है। असमिया से अंग्रेजी 9.72 और अंग्रेजी से असमिया 5.02 अनुवाद के BLEU स्कोर दर्शाया गया है। अंगजी से असमिया का अनुवाद असमिया से अंग्रेजी अनुवाद की तुलना में अपेक्षाकृत कठिन है।

5) Jawaid et al. (2014) के द्वारा अंग्रेजी से उर्दू भाषा के समानांतर कॉर्पस का उपयोग कर सांख्यिकीय आधारित एक अनुवाद प्रणाली विकसित की गयी है। इस अनुवाद प्रणाली के लिए बेसलाइन वाक्यांश आधारित और पदानुक्रमित मशीन अनुवाद का निर्माण किया गया है। दोनों अनुवाद मॉडल का आउटपुट मैन्युअल रूप से विश्लेषण किया गया है। जिनमें अंग्रेजी–से–उर्दू अनुवाद के लिए वाक्यांश–आधारित मीट्रिक टन से अधिक पदानुक्रमित मॉडल को पसंद किया जाता है। 175 वाक्यों का मैनुअल रूप से मूल्यांकन किया गया है जिसमें 45% वाक्यों में hierarchal अनुवाद प्रणाली PBMT से बेहतर परिणाम देता **है,** 21% वाक्यों में PBMT बेहतर परिणाम देता है तथा शेष वाक्यों में दोनों समान हैं।

6) Dungarwal et al. (2014) ने अंग्रेजी–हिंदी और हिंदी–अंग्रेजी सांख्यिकीय प्रणाली का निर्माण किया है। इस अनुवाद प्रणाली के मुख्य घटक वाक्यांश आधारित (हिंदी–अंग्रेजी) और फ़ैक्टर्ड (अंग्रेजी–हिंदी) मशीन अनुवाद प्रणाली है। जिसमें कारक, वचन, ट्री ऐडजाइंट व्याकरण अंग्रेजी–हिंदी को बेहतर बनाने में मदद करता है। कुछ प्राथमिक योगदान इस प्रकार हैं:

- संरचनात्मक रूप से जटिल वाक्यों के बेहतर अनुवाद के लिए सुपरटैग कारकों का उपयोग किया गया है।
- हिंदी में संज्ञा विभक्तियों को सटीक रूप से उत्पन्न करने के लिए संख्या– कारकों का उपयोग किया गया है।
- हिंदी स्रोत कॉर्पस के लिए पूर्व–अनुक्रम shallow पार्सिंग का उपयोग किया गया है।

7) Prabhugaonkar R Neha et al. (2014) ने पाँच भारतीय भाषा जोड़े जैसे बंगाली–हिंदी, अंग्रेजी–हिंदी, मराठी–हिंदी, तमिल–हिंदी और तेलुगु–हिंदी के लिए पंद्रह पदानुक्रमिक वाक्यांश आधारित सांख्यिकीय मशीन अनुवाद (HPBSMT) प्रणाली विकसित करने का प्रयास किया है। जो तीन अलग–अलग डोमेन स्वास्थ्य, पर्यटन और जनरल में है। इसका नाम पंचभूता दिया गया है। इसमें cdec टूलकिट का उपयोग करके ट्रेन, ट्यून और परीक्षण करने के लिए एक बहुत ही सरल दृष्टिकोण का उपयोग किया गया है। तथा उपयोग किए गए कॉर्पोरा का वर्णन किया है और विस्तार से प्रणालियों को प्रशिक्षित करने का विवरण दिया गया है। प्रणाली का मूल्यांकन भी किया गया है।

8) S. Sreelekha., Pushpak Bhattacharyya. (2016) ने अपने इस शोध–पत्र में मशीन अनुवाद अंग्रेजी

और मलयालम के लिए शाब्दिक संसाधनों के उपयोग कर वाक्यांश आधारित मशीन अनुवाद प्रणाली के विभिन्न सांख्यिकीय मशीन अनुवाद (SMT) प्रणालियों का तुलनात्मक वर्णन किया है। प्रशिक्षण कॉर्पस को समृद्ध करने के लिए शाब्दिक संसाधनों को दो तरीकों से संवर्धित किया गया है (A) अतिरिक्त शब्दावली और (B) विभक्त मौखिक रूप। शाब्दिक ससाधनों में इंडोवर्डनेट सिमेंटिक रिलेशन सेट, शाब्दिक शब्द और क्रिया वाक्यांश आदि शामिल हैं। सांख्यिकीय मॉडल के लिए moses और Giza ++ 2 का तकनीकी का उपयोग किया गया है। अनुवाद प्रणाली का परीक्षण health ILCI पर्यटन, स्वास्थ्य 'से लिए गए 2000 वाक्यों के कोष के साथ किया गया है। इसके अलावा 500 वाक्यों के एक ट्यूनिंग (MERT) का उपयोग किया है। व्यक्तिपरक मूल्यांकन, BLEU स्कोर, METEOR और TER जैसे विभिन्न तरीकों का उपयोग करके मलयालम अंग्रेजी और अंग्रेजी मलयालम के अनुवादित आउटपुट का मूल्यांकन किया गया है। अनुवाद की धारिता मलयालम से अंग्रेजी के मामले में 85.34% और अंग्रेजी से मलयालम के मामले में 87% तक बढ़ जाती है।

9) Lakshmikanth Mr G., Lakshmi Smt. B. Dhana. (2016) के द्वारा तेलुगु से अंग्रेजी वाक्यांश आधारित सांख्यिकीय मशीन अनुवाद प्रणाली विकसित की गई है। IRST लैंग्वेज मॉडल टूल किट (IRSTLM) लैंग्वेज मॉडल, GIZA ++ और ट्रांसलेशन मॉडल के लिए mkcls, डिकोडिंग के लिए moses का उपयोग किया गया है। तथा 43,500 से अधिक वाक्यों के एक समानांतर कॉर्पस को विकसित किया गया है जिसमें स्वतंत्रता सेनानियों के जीवन के इतिहास का छोटे–छोटे वाक्य है, जो अदालतों में उनके निशान के संदर्भ में हैं। तेलुगु और अंग्रेजी में 10760 वाक्यों के एक समानांतर कॉर्पस का उपयोग प्रणाली के प्रशिक्षण में किया गया है। 100 तेलुगु वाक्यों का अनुवाद अंग्रेजी भाषा में किया गया। मानव मूल्यांकन पद्धति का उपयोग करके 100 वाक्यों के अनुवाद का मूल्यांकन किया गया है। प्रवाह और पर्याप्तता के मापदंडों पर क्रमशः 2.693 और 2.93 की ज्यामितीय औसत गणना की गई।

10) ISLAM & PURKAYASTHA. (2018) ने अंग्रेजी से बोडो सांख्यिकीय मशीन अनुवाद प्रणाली के अनुवाद परिणाम में वृद्धि करने हेतु अंग्रेजी से बोडो मशीन अनुवाद लिप्यंतरण प्रणाली का निर्माण किया है। इस प्रणाली में पदबंध आधारित मशीन अनुवाद दृष्टिकोण मं, मोजेज़ Kenlm, N–gram तकनीक, GIZA++ और BLEW तकनीक का उपयोग किया गया है। तथा अंग्रेजी और बोडो भाषा के लिए 6000 समानांतर वाक्यों का उपयोग करके परीक्षण किया गया है। BLEU स्कोर जो अंग्रेजी से बोडो मशीन अनुवाद प्रणाली के लिए प्रत्येक डोमेन (अंग्रेजी–बोडो समानांतर पाठ कॉर्पस) के लिए अंग्रेजी में प्राप्त किया गया है जो नीचे तालिका में दिखाया गया है।

Multi–domain English–Bodo Parallel Text Corpora	Corpus Statistics (Sentences)			BLEU scores	
	Training	Tuning	Testing	Before using the MTn System	After using the MTn System
Agriculture	3500	500	3500	30.18	31.92
Health	12000	1000	12000	38.87	40.08
Tourism	9000	1000	9000	37.50	38.35

11) pandey et al. (2018) ने एक वाक्यांश आधारित संस्कृत–हिंदी (SaHiT) सांख्यिकीय मशीन अनुवाद प्रणाली प्रस्तुत किया है। इस मशीन अनुवाद प्रणाली के लिए मोजेज़ टूलकिट का प्रयोग किया गया है। संस्कृत–हिंदी समानांतर कॉर्पस के 43k वाक्यों और लक्ष्य भाषा (हिंदी) में एक एकभाषी कॉर्पस के 56k वाक्यों का उपयोग किया गया है। यह प्रणाली 57 BLEU स्कोर देता है। इस संस्कृत– हिंदी मशीनी

अनुवाद प्रणाली का मूल्यांकन तीन मूल्यांकनकर्ताओं द्वारा किया गया था। उन्होंने पर्याप्तता और प्रवाह के आधार पर मशीन अनुवाद के आउटपुट को आंका गया है। मूल्यांकनकर्ताओं द्वारा दिए गए 1–5 के बोच स्कोर के आधार पर पर्याप्तता और प्रवाह की गणना की जाती है। 91% पर्याप्तता और 66.72% प्रवाह प्राप्त किया गया।

12) Daimary, Maheswar et al. (2019) द्वार इस प्रस्तावित शोध कार्य का मुख्य उद्देश्य बोडो–अंग्रेजी पदबंध आधारित मशीन अनुवाद प्रणाली को विकसित करना है, यहां एक सांख्यिकीय मशीन अनुवाद इंजन Moses का उपयोग स्रोत भाषा से लक्ष्य भाषा में पाठ अनुवाद के सांख्यिकीय मॉडल को प्रशिक्षित करने के लिए किया गया है। शब्दों को संरेखित करने के लिए भाषा मॉडल और GIZA ++ उपकरण को विकसित करने के लिए IRSTLM टूल का उपयोग किया गया है। इसमें क्रमशः 8000 बोडो और अंग्रेजी समानांतर वाक्यों के साथ प्रणाली को प्रशिक्षित किया गया है। बोडो–अंग्रेजी मशीन अनुवाद प्रणाली की कई बार यात्रा और पर्यटन से संबंधित समानांतर कॉर्पोरा का व्यक्तिगत रूप से की जांच की गई है। यह देखा गया है कि प्रत्येक डोमेन समानांतर कॉर्पस में समानांतर वाक्यों की संख्या बढ़ाकर अनुवाद परिणाम को बढ़ाया जा सकता है। जब भी कॉर्पस की मात्रा बढ़ाया जाता है, तो अनुवाद की गुणवत्ता में भी सुधार होता है।

V. मशीन अनुवाद प्रणालियों का विवरण:

S- No-	Title of the Research Paper	Year	Approach	Per- ¼%½
1-	English to Bangla Phrase&Based Machine Translation	2010	Statistical based machine translation	11-7
2-	Modeling Phrase&Based Statistical Machine Translation for English&Hindi Language	2012	Statistical based machine translation	&&&&&&&&&&
3-	Making Headlines in Hindi: Automatic English to Hindi News Headline Translation	2013	phrase&based and factor&based	13-40 and 5-73
4-	ASSAMESE&ENGLISH BILINGUAL MACHINE TRANSLATION	2014	statistical phrase based translation	9-72 and 5-02
5-	English to Urdu Statistical Machine Translation: Establishing a Baseline	2014	statistical phrase based translation	&&&&&&&&&&&&&&
6-	The IIT Bombay Hindi]English Translation System at WMT 2014	2014	statistical phrase based translation	&&&&&&&&&
7-	PanchBhoota: Hierarchical Phrase Based Machine Translation Systems for Five Indian Languages	2014	statistical phrase based translation	&&&&&&&&&&&&
8-	Lexical Resources to Enrich English Malayalam Machine Translation	2016	statistical phrase based translation	85-34% and 87%
9-	An Approach for Telugu to English Phrase Based Statistical Machine Translation System	2016	statistical phrase based translation	2-693 and 2-93
10-	English to Bodo Machine	2017	statistical	31-92] 40-08

	Transliteration System for Statistical Machine Translation		phrase based translation	and 38-35
11-	Demo of Sanskrit&Hindi SMT System	2018	statistical phrase based translation	91% and 66-72
12-	Bodo To English Statistical Machine Translation System	2019	statistical phrase based translation	&&&&&&&& &&&&

VI. निष्कर्ष:

उपरोक्त समीक्षा से यह निष्कर्ष निकलता है कि प्राकृतिक भाषा संसाधन के क्षेत्र में बहुत सारे शोध किए जा रहे हैं और प्रत्येक दृष्टिकोण में भाषा युग्मों को कुछ सीमा के साथ अनुवाद करने की क्षमता है। प्रस्तुत शोध–पत्र में मशीन अनुवाद प्रणाली के निर्माण से संबंधित पूर्व में किए गए शोधकार्यों तथा उनमें प्रयुक्त पदबंध आधारित सांख्यिकीय मशीनी अनुवाद तकनीकों का संक्षिप्त विवरण प्रस्तुत किया गया है। भारतीय भाषाओं के लिए विकसित मशीन अनुवाद प्रणालियों में से अधिकांश ने सांख्यिकीय दृष्टिकोण का पालन किया है। कारण यह है कि भारतीय भाषाएं सुविधाओं में बहुत अधिक समृद्ध हैं और प्रकृति में कृषि प्रधान हैं, इसलिए पूर्णतया मशीन अनुवाद प्रणाली विकसित करने के लिए नियम–आधारित दृष्टिकोण कई स्थितियों में विफल रहे हैं। सांख्यिकीय दृष्टिकोण ने शोधकर्ताओं को भारतीय भाषाओं के लिए मशीन अनुवाद प्रणाली विकसित करने के लिए इन तरीकों को चुनने के लिए प्रोत्साहित किया है। जिससे मेरे शोध कानूनी भाषा हेतु अंग्रेजी–हिंदी मशीन अनुवाद प्रणाली हेतु उचित एवं प्रभावी प्रविधि निर्माण में सहायता मिलेगी।

VII. संदर्भ–सूची

- Koehn, Philipp., Och, Franz J., & Marcu, Daniel. (2003). statistical Phrase-Based Translation. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics , 127–133.
- Islam, Md. Zahurul., Tiedemann, J¨org., & Eisele, Andreas. (2010). English to Bangla Phrase-Based Machine Translation. 1-8.
- Singh, Prof. Dharvendra., Agrawal, Rajul., Dalal, Ritu. (2012). Modelling Phrase-Based Statistical Machine Translation for English-Hindi Language. IJRREST: International Journal of Research Review in Engineering Science and Technology (ISSN 2278- 6643)|Volume-1 Issue-3, , 95-99.
- Joshi, Aditya., Popat, Kashyap., Gautam, Shubham., Bhattacharyya, Pushpak. (2013). Making Headlines in Hindi: Automatic English to Hindi. The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations, , 21-24.
- Baruah, Kalyanee Kanchan., Das., Pranjal., Hannan, Abdul., & Sarma, Shikhar Kr. (2014). ASSAMESE-ENGLISH BILINGUAL MACHINE TRANSLATION. International Journal on Natural Language Computing (IJNLC) Vol. **3**, No.**3**, , 73-82.
- Jawaid, Bushra., Kamran, Amir., & Bojar, Ondřej. (2014). English to Urdu Statistical Machine Translation: Establishing a Baseline. Proceedings of the **5**th Workshop on South and Southeast Asian NLP, **25**th International Conference on Computational Linguistics, 37-42.
- Dungarwal, Piyush., Chatterjee, Rajen., Mishra, Abhijit., Kunchukuttan, Anoop., Shah, Ritesh., & Bhattacharyya, Pushpak. (2014). The IIT Bombay Hindi, English Translation System at WMT 2014. Proceedings of the Ninth Workshop on Statistical Machine Translation, , 90–96.
- Prabhugaonkar, Neha R., Pawar, Jyoti., Nagvenkar, Apurva S., Bhattacharyya, Pushpak., Kanojia, Diptesh., & Shrivastava, Manish. (2014). PanchBhoota: Hierarchical Phrase Based

Machine Translation. Conference: SMT Contest in International Conference on Natural Language Processing (ICON 2014), At Goa, India, Volume: Eleventh, 1-6.

- S, Sreelekha., & Bhattacharyya, Pushpak. (2016). Lexical Resources to Enrich English Malayalam Machine Translation. 10th edition of the Language Resources and Evaluation Conference, , 620–627.
- Lakshmikanth, Mr G., & Lakshmi, Smt.B.Dhana. (2016). An Approach for Telugu to English Phrase Based Statistical Machine Translation System. International Jouranl & Magazine of Engineering, Technology, Management and Research A Peer Reviewed access International Jouranl Volume No. 3 , 617-627.
- Singh, Avinash., Kour, Asmeet, & Shubhnandan, S. Jamwal. (2016). English-Dogri Translation System using MOSES. Jamwal Department of Computer Science & IT University of Jammu.
- Islam, Saiful., & Purkayastha, Bipul Syam. (2018). English to Bodo Machine Transliteration System for Statistical Machine Translation. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, , 7989-7997.
- Pandey, Rajneesh., Ojha, Atul Kr., & Jha, Girish Nath. (2018). Demo of Sanskrit-Hindi SMT System. http://lrec-conf.org/workshops/lrec2018/W11/summaries/20_W11.html, 1-2.
- Daimary, Maheswar., Sarma, Shikhar Kumar., & Rahman, Mirzanur. (2019). Bodo To English Statistical Machine Translation System. International Journal of Computer Sciences and Engineering , 1731-1736.