भारतीय वैज्ञानिक एवं औद्योगिक अनुसंधान पत्रिका
वर्ष 29 अंक (1) जून 2021 पृ. 49-54

मशीन आधारित भाषा अनुवाद में संदर्भ निराकरण का महत्व

चेतन अग्रवाल एवं कमलेश दत्ता
संगणक विज्ञान एवं अभियान्त्रिकी विभाग
राष्ट्रीय प्रौद्योगिकी संस्थान, हमीरपुर 177 005 (हिमाचल प्रदेश)

सारांश : निरंतर विकासशील सूचना प्रौद्योगिकी के इस युग में विभिन्न प्रकार की सूचना व जानकारी का बड़े व्यापक स्तर पर निरंतर आदान-प्रदान होता रहता है। परन्तु सूचना व ज्ञान के इस भण्डार का विश्व की विभिन्न भाषाओं में उपलब्ध होना इसके समुचित उपयोग में कुछ अवरोध उत्पन्न करता है, जिसे त्वरित भाषा अनुवाद के माध्यम से कुछ सीमा तक दूर किया जा सकता है। मशीन आधारित भाषा अनुवाद इस दिशा में एक महत्वपूर्ण एवं कारगर पड़ाव साबित हुआ है। विगत दो दशकों में इन्टरनेट की आसान पहुँच के कारण मशीन आधारित भाषा अनुवाद की आवश्यकता दिन-प्रतिदिन बढ़ रही है। परन्तु मशीन आधरित भाषा अनुवाद में कई तरह की तकनीकी बाधाएँ एवं समस्याएँ आती हैं जिनका समाधान गुणवत्तापूर्ण अनुवाद हेतु आवश्यक हो जाता है। पूर्व उल्लेखित शब्द-समूह संदर्भ निराकरण उनमें से एक प्रमुख एवं जटिल समस्या है, जिसने निरंतर मशीन आधारित भाषा अनुवाद के क्षेत्र में कार्यरत कई शोधकर्ताओं एवं शोध-प्रतिष्ठानों का ध्यान आकर्षित किया है। यद्यपि लगभग सभी भाषाओं में इस तरह के शब्द-समूह का उपयोग होता है, परन्तु भाषानुसार इनका प्रकार अलग-अलग हो सकता है। अतः, स्त्रोत भाषा व प्रस्तावित अनुवादित भाषा के आधार पर पूर्व उल्लेखित शब्द-समूह के संदर्भ निराकरण का प्रकार भी भिन्न हो सकता है। प्रस्तुत आलेख विभिन्न मशीन आधारित भाषा अनुवाद प्रारूप में उपस्थित अलग-अलग प्रकार के संदर्भ-बोधक, उनके निराकरण से सम्बंधित सैद्धांतिक तरीके एवं एल्गोरिद्म के विस्तृत अवलोकन का एक प्रयास है।

Importance of anaphora resolution in machine translation

Chetan Agarwal & Kamlesh Dutta
Department of Computer Science and Engineering
National Institute of Technology, Hamirpur 177 005 (Himachal Pradesh)

Abstract

In this era of continuously progressing Information Technology, information exchange is taking place in various domains at a faster pace. However, availability of huge information, diversity of knowledge sources and variations in structure of languages, leads to some hurdles in its effective utilization. Use of Machine Translation is a good enough solution at this stage, which can remove several constraints up-to some extent. Also, because of the easy and economic Internet accessibility, Machine Translation is becoming crucial. But several technical constraints are there in Machine Translation process that must be resolved for qualitative translation. Anaphora resolution is one of the major and complex problems of Machine Translation. Many researchers and research institute working in the field of language technology and translation are continuously striving to resolve the problem. Depending upon type of language there may be various classes of anaphor.Therefore, depending upon the type of source and target language, type of anaphor and anaphor resolution process also may differ. This paper is an attempt to go in detail and reviewing of currently available theoretical approaches and algorithms of anaphora resolution that are being used in various machine translation models.

प्रस्तावना

सामान्यतः सभी प्रकार के लेखन-कार्य अथवा संभाषण में रुचि व शब्दों की विविधता बनाये रखने के लिए, किसी शब्द या शब्द-समूह विशेष की पुनरावृत्ति से बचने के लिए, प्रायः किसी सर्वनाम शब्द या अन्य शब्द का उपयोग किया जाता है। इस तरह किसी वाक्य में पूर्व-उल्लेखित शब्द-समूह के स्थान पर प्रयुक्त सर्वनाम या अन्य शब्द-समूह को संदर्भ-बोधक कहते हैं, अर्थात् संदर्भ-बोधक वह शब्द या शब्द-समूह है जो किसी पूर्व-उल्लेखित

शब्द-समूह को इंगित करते है[1]। इस शब्द की व्युत्पत्ति प्राचीन ग्रीक-शब्द Ana ,oa Phora के संयोजन से हुई है। Ana का अर्थ है ''पूर्व-वर्णित'' एवं Phor का अर्थ है ''इंगित करना''[2&3]। इस प्रकार Anaphora का अर्थ है पूर्व-वर्णित शब्द-समूह के संदर्भ को इंगित करना। यद्यपि इस प्रकार विविध शब्दों के उपयोग से लेखन या वार्ता में रोचकता तो बनी रहती है, परन्तु कभी-कभी उचित संदर्भ निराकरण नहीं होने की स्थिति में इन शब्दों के प्रयोग से स्रोता या पाठक को उचित संदर्भ समझने में कुछ कठिनाई भी हो जाती है। मशीन आधारित भाषा अनुवाद के लिए संदर्भ निराकरण एक प्रमुख एवं जटिल समस्या है जिसके कारण मशीन आधारित भाषा अनुवाद की गुणवत्ता प्रभावित होती है[4-6]। इसे निम्न उदाहरण की सहायता से समझ सकते हैं।

उदाहरण 1.

1. We purchased a house in outskirt of city because it is a peaceful place.
2. We purchased a house in outskirt of city because there it is cheap.

उपरोक्त वाक्य सं. 1 व 2 में "it" शब्द क्रमशः ''outskirt of city'' और ''a house'' के लिए प्रयोग किया गया है, जबकि वाक्य सं. 2 में "there" शब्द का प्रयोग "outskirt of city" के स्थान पर किया गया है। यदि इस "it" और "there" के संदर्भ का सही प्रकार से निराकरण नहीं होता है तो अंग्रेजी से हिंदी में अनुवाद करते समय अनुवाद गलत होने की सम्भावना अधिक हो जाती है। जैसे इस उदाहरण के लिए यदि वाक्य सं. 2 में "it" और "there" शब्द को क्रमशः "outskirt of city" व "a house" के स्थान पर प्रयुक्त हुआ मानते हैं तो इस वाक्य के सही हिंदी अनुवाद ''हमने शहर से दूर घर खरीदा है, क्योंकि वहाँ पर यह सस्ता है।'', की जगह पर इसका गलत अनुवाद ''हमने शहर से दूर घर खरीदा है, क्योंकि यह वहां पर सस्ता है।'' हो सकता है, जो "outskirt of city is cheap in house" का अनुवाद होता। इस प्रकार, पूर्व उल्लेखित शब्द-समूह संदर्भ निराकरण की भाषा अनुवाद में प्रासंगिकता दो महत्वपूर्ण तथ्यों को रेखांकित करती है; पहला मशीन आधारित भाषा अनुवाद में संदर्भ की अस्पष्टता का निवारण व दूसरा वाक्यशः अनुवाद के स्थान पर समेकित अनुवाद की भूमिका।

प्रस्तुत आलेख के भाग-2 में विभिन्न प्रकार के संदर्भ-बोधक शब्द-समूह का वर्णन है एवं भाग-3 में संदर्भ निराकरण के मूलभूत सिद्धांतों के बारे में बताया गया है। भाग-4 संदर्भ निराकरण की कुछ महत्वपूर्ण एल्गोरिद्म का एक अवलोकन है।

संदर्भ बोधक के प्रकार व संदर्भ निराकरण से सम्बंधित समस्याएँ

यद्यपि विभिन्न भाषा व्याकरण के आधार पर संदर्भ-बोधक शब्द-समूह भी विभिन्न प्रकार के हो सकते हैं, परन्तु उनमें से कुछ संदर्भ-बोधक शब्द-समूह, जो प्रायः सभी भाषा व्याकरण में पाए जाते हैं, केवल उनका ही संक्षिप्त वर्णन प्रस्तुत आलेख में किया गया है। कुछ शोधकत्ताओं ने संदर्भ बोधक शब्द-समूह के प्रकार एवं संदर्भ निराकरण से सम्बंधित समस्याओं का विस्तार से वर्णन किया है।

• सर्वनाम मूलक संदर्भ

सर्वनाम मूलक संदर्भ-शब्द वो सर्वनाम होते हैं जो किसी पूर्व उल्लेखित शब्द-समूह के संदर्भ का बोध कराते हैं।

उदाहरण 2. गणतंत्र के शुभ अवसर पर राष्ट्रपति महोदय ने राष्ट्र के नाम अपने संबोधन में लोकतंत्र की विशेषताओं का जिक्र किया। उन्होंने लोकतंत्र में लोक व तंत्र की परस्पर पूरकता के विषय में अपने विचार व्यक्त किये।

उपरोक्त उदाहरण के दूसरे वाक्य में ''उन्होंने'' शब्द एक सर्वनाम शब्द है जो पहले वाक्य में वर्णित संज्ञा-शब्द समूह ''राष्ट्रपति महोदय'' के संदर्भ को इंगित करता है।

• निश्चित संज्ञा मूलक संदर्भ

निश्चित संज्ञा मूलक संदर्भ शब्द वो संज्ञा मूलक शब्द होते हैं जो किसी पूर्व उल्लेखित संज्ञा शब्द-समूह के स्थान पर प्रयोग होते हैं।

उदाहरण 3.

गणतंत्र के शुभ अवसर पर राष्ट्रपति महोदय ने राष्ट्र के नाम अपने संबोधन में लोकतंत्र की विशेषताओं का जिक्र किया। आदरणीय महामहिम ने इस अवसर पर परेड का निरीक्षण भी किया।

उपरोक्त उदाहरण के दूसरे वाक्य में ''आदरणीय महामहिम'' शब्द एक संज्ञा मूलक शब्द-समूह है जो पहले वाक्य में वर्णित संज्ञा-शब्द समूह ''राष्ट्रपति महोदय'' के स्थान पर प्रयुक्त किया गया है।

• एकाकी संज्ञा मूलक संदर्भ

एकाकी संज्ञा मूलक संदर्भ शब्द वे संदर्भ-शब्द होते हैं जो समान प्रकार के एक से ज्यादा संज्ञा-सूचक शब्दों में से किसी एक विशेष संज्ञा-सूचक शब्द को इंगित करने के लिए प्रयोग किये जाते हैं।

उदाहरण 4.

गणतंत्र-दिवस की पूर्व-संध्या पर लालकिले पर एक कवि-सम्मलेन का भी आयोजन किया गया। इसका सीधा प्रसारण दूरदर्शन-1 और दूरदर्शन-2, दोनों चैनलों पर किया गया। जो लोग पहले वाले चैनल पर नहीं देख पाए वे दूसरे वाले चैनल पर इसका आनंद ले सकते थे।

उपरोक्त उदाहरण के दूसरे वाक्य में ''पहले वाले'' व ''दूसरे वाले'' एकाकी संज्ञा मूलक संदर्भ-शब्द हैं जो पहले वाक्य में वर्णित संज्ञा-शब्द समूह ''दूरदर्शन-1'' और ''दूरदर्शन-2'' के स्थान पर प्रयुक्त हुए हैं।

• संख्या मूलक संदर्भ

संख्या मूलक संदर्भ-शब्द सामान्यतः समान बोध वाले पूर्व उल्लेखित शब्द-समूहों की संख्या या गिनती के बोध के लिए प्रयोग होते हैं।

उदाहरण 5.

गणतंत्र दिवस और स्वतंत्रता दिवस, दोनो भारत के राष्ट्रीय पर्व हैं।

उपरोक्त उदाहरण में ''दोनों'' शब्द एक संख्या मूलक शब्द-समूह है जो पहले वर्णित ''गणतंत्र दिवस और स्वतंत्रता दिवस ''संज्ञा-शब्द समूह के संदर्भ में प्रयुक्त किया गया है।

• अनिश्चित संज्ञा मूलक संदर्भ

अनिश्चित संज्ञा मूलक संदर्भ शब्दों का प्रयोग ऐसे पूर्व उल्लेखित शब्द-समूह के लिए किया जाता है जो व्यवहारिक रूप में अनिश्चित हो अर्थात् संदर्भ बोधक शब्द किसी निर्दिष्ट संज्ञा के लिए प्रयुक्त हुआ हो।

उदाहरण 6.

गणतंत्र दिवस के अवसर पर सेना के जवानों ने कई तरह के हैरत-अंगेज करतब दिखाए, जिसके लिए राष्ट्रपति ने उनकी खूब प्रशंसा की।

उपरोक्त उदाहरण में ''सेना के जवानों'' शब्द-समूह व्यवहार में अनिश्चित प्रकार का संज्ञा मूलक शब्द-समूह है, अतः ''उनकी'' शब्द जो ''सेना के जवानों'' के सन्दर्भ में प्रयोग किया गया है, वह एक समूह विशेष को निरूपित करता है, न कि किसी निर्दिष्ट संज्ञा को।

• संकेत मूलक संदर्भ

संकेत मूलक संदर्भ शब्दों का प्रयोग पूर्व वर्णित संज्ञा-शब्द समूह के स्थान पर तृतीय-पुरुष के भाव में किया जाता है।

उदाहरण 7.

गणतंत्र दिवस के अवसर पर ''नव-निर्मित युद्धपोत INS विक्रमादित्य'' को भी प्रदर्शन के लिया रखा गया। यह एक अति-विशाल और सभी आधुनिक सुविधाओं से युक्त युद्धपोत है।

उपरोक्त उदाहरण में ''नव-निर्मित युद्धपोत INS विक्रमादित्य'' शब्द-समूह एक निश्चित प्रकार का संज्ञा मूलक शब्द-समूह है एवं ''यह'' शब्द पूर्व वर्णित विशेष प्रकार के युद्धपोत के संदर्भ में प्रयोग किया गया है, अर्थात् किसी निर्दिष्ट संज्ञा को इंगित करता है।

संदर्भ निराकरण के सिद्धान्त एवं प्रक्रिया

यद्यपि वैसे तो संदर्भ-बोधक शब्द किसी संज्ञा-सूचक वाक्यांश अथवा क्रिया-सूचक वाक्यांश या वाक्य- समूह अथवा परिच्छेद इत्यादि किसी के भी संदर्भ में उपयोग किये जा सकते हैं, परन्तु अधिकांश संदर्भ निराकरण प्रक्रिया केवल संज्ञा-सूचक वाक्यांश के लिए ही कार्यक्षम है। प्रायः किसी भी लेखन या संभाषण में उपस्थित संदर्भ बोधक शब्दों से सम्बंधित पूर्व-उल्लेखित शब्द-समूह की उपस्थिति, उसी वाक्य में या ज्यादा से ज्यादा दो या तीन वाक्य पहले होती है, परन्तु कभी-कभी इन वाक्यांशों की उपस्थिति संबंधित संदर्भ बोधक शब्द से 8-10 वाक्य पहले भी हो सकती है। इसलिए किसी लेखन या संभाषण में जब भी कोई संदर्भ बोधक शब्द आता है तो संदर्भ निराकरण प्रक्रिया में सर्वप्रथम संभावित पूर्व-उल्लेखित शब्द-समूह की उपस्थिति क्षेत्र को चिन्हित करना आवश्यक है। तत्पश्चात् चिन्हित वाक्य-समूह में सभी संभावित पूर्व-उल्लेखित शब्द-समूहों को उम्मीदवार मानते हुए, व्याकरण, विषय, लिंग, वचन, पुरुष, संख्या इत्यादि के आधार पर कुछ पूर्व-उल्लेखित शब्द-समूहों को रद्द करते हुए, बचे हुए उम्मीदवारों में से वरीयता के आधार पर चयन किया जाता है। इस प्रकार संदर्भ निराकरण प्रक्रिया को दो मुख्य चरणों में पूरा किया जाता है-

1. संभावित पूर्व उल्लेखित शब्द-समूह (Possible Antecedent) की पहचान व चयन।
2. व्याकरण संबंध व अन्य भाषा-बोध के आधार पर वरीयता या निरस्तीकरण।

इस प्रकार, संदर्भ निराकरण प्रक्रिया के दूसरे चरण में वरीयता व निरस्तीकरण दो ऐसे प्रमुख कारक हैं जिनके आधार पर किसी भी संदर्भ-बोधक शब्द के लिए उचित पूर्व-उल्लेखित शब्द-समूह का निर्धारण किया जाता है। निम्न उदाहरणों द्वारा दोनों ही प्रकार के कारकों की निराकरण प्रक्रिया में भूमिका को समझ सकते हैंः-

• **निरस्तीकरण कारक**

• **लिंग व वचन बंधन**

लिंग व वचन बंधन के अनुसार किसी भी संज्ञा-सूचक शब्द-समूह व उसके लिए प्रयोग किये गए संदर्भ-बोधक शब्द के लिंग व वचन में समानता होनी चाहिए।

उदाहरण 8.

द्रोणाचार्य व अन्य प्रमुख कौरव महारथियों ने अभिमन्यु को चक्रव्यूह में घेरने की योजना बनाई। परन्तु उसने उन सभी को परास्त कर दिया।

उपरोक्त उदाहरण में दो संदर्भ-बोधक शब्द, उसने व उन सभी का प्रयोग हुआ है जो क्रमशः अभिमन्यु व द्रोणाचार्य व अन्य प्रमुख कौरव महारथियों के संदर्भ में है। इस उदाहरण में अभिमन्यु शब्द एकवचन है जबकि द्रोणाचार्य व अन्य प्रमुख कौरव महारथियों बहुवचन है, तथा उसने एकवचन संदर्भ-बोधक व उन सभी बहुवचन संदर्भ-बोधक है। इस प्रकार वचन के आधार पर संदर्भ निराकरण किया गया है।

• **वाक्य रचना सम्बन्धित नियम**

उपरोक्त उदाहरण स. 8 में यद्यपि उसने संदर्भ-बोधक द्रोणाचार्य व अभिमन्यु दोनों ही संज्ञा-शब्द के संदर्भ में हो सकता है, क्योंकि द्रोणाचार्य व अभिमन्यु दोनों ही एक वचन है, परन्तु हिंदी वाक्य विन्यास के अनुसार वाक्य ''द्रोणाचार्य व अन्य प्रमुख कौरव महारथियों ने अभिमन्यु को चक्रव्यूह में घेरने की योजना बनाई'' में वाक्यांश ''द्रोणाचार्य व अन्य प्रमुख कौरव महारथियों'' क्रिया ''चक्रव्यूह में घेरने की योजना'' के लिए कर्ता के रूप में है, जबकि अभिमन्यु इस क्रिया के लिए कर्म के रूप में है। इसी तरह वाक्य ''परन्तु उसने उन सभी को परास्त कर दिया'' में संदर्भ-बोधक उन-सभी वाक्यांश ''द्रोणाचार्य व अन्य प्रमुख कौरव महारथियों'' के संदर्भ में है अथार्थ ''द्रोणाचार्य व अन्य प्रमुख कौरव महारथियों'' के संदर्भ में है। इस प्रकार वाक्य-विन्यास के आधार पर भी संदर्भ निराकरण किया जा सकता है।

• **अर्थ व प्रकृति समरूपता**

इस नियम के अनुसार शब्द या शब्द-समूह की प्रकृति अथवा शब्द-रूप के आधार पर भी संदर्भ निराकरण किया जा सकता है।

उदाहरण 9.

कृष्ण ने अर्जुन को कर्म का उपदेश दिया और उसने उस पर अमल भी किया।

उपरोक्त वाक्य में उसने व उस संदर्भ-बोधक शब्दों का प्रयोग क्रमशः ''अर्जुन'' व ''कर्म का उपदेश'' के लिए किया गया है। इस उदाहरण में ''कर्म का उपदेश'' वाक्यांश क्रिया कारक शब्द ''दिया'' के कर्म-रूप में है, एवं संदर्भ-बोधक शब्द ''उस'' भी कर्म-रूप में है, इसलिए शब्द के अर्थ व प्रकृति के आधार पर ''कृष्ण'' व ''अर्जुन'' को ''उस'' संदर्भ-बोधक के लिए अमान्य किया गया है।

• **वरीयता के आधार पर**

यदि किसी संदर्भ-बोधक शब्द के लिए कई सारे संभावित पूर्व-उल्लेखित शब्द-समूह उपस्थित हों एवं उनमें से किसी को भी उपरोक्त वर्णित निरस्तीकरण कारक द्वारा निरस्त नहीं किया जा सकता हो तो इस स्थिति में शब्द-समूह की समरूपता या वाक्य-रचना में शब्द-समूह की भूमिका के आधार पर एक शब्द-समूह विशेष को दूसरे पर वरीयता दी जा सकती है।

इसे निम्न उदाहरणों से समझ सकते हैं :-

• **सरंचनात्मक समानता**

उदाहरण 10.

डॉ. कलाम ने पहले अग्नि के साथ नाग मिसाइल का परीक्षण किया। परन्तु बाद में उन्होंने इसके साथ पृथ्वी मिसाइल का परीक्षण किया।

उपरोक्त उदाहरण में उन्होंने व इसके, दो संदर्भ-बोधक शब्द हैं, जिसमें से प्रथम संदर्भ-बोधक उन्होंने के लिए संदर्भ निराकरण वाक्य-रचना संबंधी नियम से किया जा सकता है। परन्तु दूसरे संदर्भ-बोधक इसके के लिए बचे हुए दो संदर्भ अग्नि व नाग में से किसी को भी निरस्तीकरण नियम से निरस्त नहीं किया जा सकता। इस प्रकार की स्थिति में निरस्तीकरण की जगह संरचना समानता के आधार पर वरीयता दे कर समाधान कर सकते हैं। जैसे पहले वाक्य में वाक्यांश ''अग्नि के साथ नाग'' व दूसरे वाक्य में वाक्यांश ''इसके साथ पृथ्वी'' में सरंचनात्मक समानता है, इसलिए इसके संदर्भ-बोधक के लिए नाग के स्थान पर अग्नि को वरीयता देकर संदर्भ निराकरण किया गया है।

संदर्भ निराकरण की एल्गोरिद्म

जैसा कि इस लेख के प्रारंभ में ही स्पष्ट किया गया है कि मशीन-आधारित भाषा अनुवाद के लिए सन्दर्भ-निराकारण एक अत्यंत जटिल व महत्वपूर्ण समस्या है जिसका समाधान गुणवत्तापूर्ण अनुवाद के लिए अत्यंत जरूरी है। इसलिए मशीन आधारित भाषा अनुवाद के क्षेत्र में कार्यरत शोधार्थियों ने भी बहुत समय पहले से ही इसके समाधान हेतु प्रयास किये हैं। कार्य-विधि एवं कार्य-पद्धति के आधार पर इन एल्गोरिद्म्स को मोटे तौर पर दो वर्गों 1. ''पारंपरिक अथवा भाषा बोध पर आधारित'' 2. ''वैकल्पिक

अथवा सांख्यिकी आधारित'' में विभाजित किया गया है। प्रस्तुत आलेख के इस भाग में उनमें से कुछ वैकल्पिक अथवा सांख्यिकी आधारित महत्वपूर्ण एल्गोरिद्म्स का यहाँ पर वर्णन किया गया है।

• Statistical/corpus Processing Approach

विभिन्न शोधपत्रों के अवलोकन से ज्ञात होता है कि संदर्भ निराकरण के लिए सांख्यिकी आधारित पद्धति के उपयोग का प्रथम प्रयास Dagan and Itai ने 1990 में किया[3,9&15]। शोधकर्ताओं ने बड़े आकार के वाक्य-कोष के लिए समान संदर्भ में प्रयोग किये गए पैटर्न्स के आकंड़े एकत्रित करने के लिए स्वचालित मॉडल प्रस्तुत किया है। सांख्यिकी आधारित इस मॉडल के अनुसार Corp में उपलब्ध समान संदर्भ में उल्लेखित पैटर्न्स का संभावित उम्मीदवार के तौर पर चयन किया जाता है। अस्पष्टता अर्थात् एक से अधिक विकल्प की स्थिति में जिस पैटर्न की आवृत्ति अधिक होती है, उसी को सही मान लिया जाता है। इस पद्धति का उपयोग करते हुए उन्होंने English pronoun "it" पर संदर्भ निराकरण संबंधी एक प्रयोग किया। अपने इस प्रयोग के परिणाम में उन्होंने पाया कि इस मॉडल के उपयोग से जो आकंड़े एकत्रित हुए वे वास्तव में Semantic Constraint से बेहद समानता रखते हैं।

• knowledge-Independent Approach

संदर्भ निराकरण में सांख्यिकी आधारित पद्धति के प्रयोग का एक अन्य महत्वपूर्ण प्रयास Nasukawa ने किया है। अपने शोध पत्र[11] में लेखक ने भाषा व व्याकरण की बिना किसी विशेष जानकारी के English Pronoun संबंधी संदर्भ निराकरण के लिए, परस्पर वाक्य-समूह में निहित सूचना (Inter&Sentential Information) का उपयोग करते हुए, 90% तक सफल, एक सटीक व सरल अल्गोरिथम प्रस्तुत करने का दावा किया है। इस शोध-पत्र में लेखक ने संदर्भ-निराकरण हेतु तीन प्रभावी कारकों का उल्लेख किया है, जो संभावित संज्ञा-सूचक वाक्यांशों में से उपयुक्त वाक्यांश का चयन करने में अतिउपयोगी है। ये तीन कारक निम्न प्रकार है;

I. Collocation patterns within asource text
II. Frequency of repetition in precedingsen-tences
III. Syntactic position

• An Uuncertainty-Reasoning Approach

Mitkov ने अपने शोधपत्र में अनिश्चित कारकों पर आधारित Artificial Intelligence Aapproach का प्रतिपादन किया। इस पद्धति के प्रतिपादन के समर्थन में ने निम्न तर्क दिए है[12];

- सामान्यतया किसी भी प्राकृतिक भाषा को समझने के लिए बनाये गए सभी कम्प्यूटर प्रोग्राम को संदर्भ-बोधक शब्दों से सम्बंधित पूर्व-उल्लेखित शब्द-समूह का चयन प्रायः आधी-अधूरी जानकारी के आधार पर करना होता है। प्रायः निरस्तीकरण कारक व वरीयता कारक की जानकारी उपलब्ध होते हुए भी कम्प्यूटर प्रोग्राम उसका उपयोग करने में सक्षम नहीं होते हैं।
- चूंकि प्रारंभ में किसी भी निरस्तीकरण कारक या वरीयता कारक का स्कोर मानव द्वारा ही निधारित किया जाता है, वे मूल-रूप में कर्ता रूप में ही हैं, अतः उन्हें भी अनिश्चित तथ्य ही मानना चाहिए।

इस एप्रोच में प्रमुख विचार यह है कि किसी भी पूर्व-उल्लेखित शब्द-समूह की खोज को परिकल्पना ''एक निश्चित संज्ञा-सूचक वाक्यांश ही सही पूर्व उल्लेखित शब्द-समूह है'' के आधार पर ही स्वीकृत या अस्वीकृत किया जाये।

• Modelling Pronominal Anaphora in Sta-tistical Machine Translation

कुछ शोधकत्ताओं ने सर्वनाम मूलक संदर्भ निराकरण के लिए, सांख्यिकी आधारित, English से German में अनुवाद हेतु, एक नया मॉडल "Word Dependency Modal" प्रतिपादित किया जो सम्बंधित पूर्व-उल्लेखित शब्द-समूह (Antecedent) की उपस्थिति की पहचान उस वाक्य या वाक्य से पूर्व भी कर सकता है[13]। किसी वाक्य में उपस्थित सर्वनाम व उससे सम्बंधित सही पूर्व-उल्लेखित शब्द-समूह का युग्म बनाने के लिए इस मॉडल में एक Open Source Co-Reference Resolution SystemBART का उपयोग किया है[14]।

• Hybrid Approach to Pronominal Anaph-ora Resolution

शोधकर्ता kamune and Agrawal ने अपने शोधपत्र में सर्वनाम मूलक संदर्भ निराकरण हेतु वरीयता कारक व निरस्तीकरण कारक आधारित आर्किटेक्चर की एक मिश्रित पद्धति को विकसित करने का दावा किया है[15]। शोधपत्र के दावे के अनुसार यह पद्धति तृतीय पुरूष सर्वनाम मूलक संदर्भ बोधक शब्द के लिए, उसी वाक्य या उस वाक्य से अधिकतम तीन-चार वाक्य पहले संभावित पूर्व-उल्लेखित शब्द-समूह की पहचान करने में सक्षम है। Java में क्रियान्वित व Charniak Parser को एक Associated tool की भांति उपयोग करने वाले इस सिस्टम की दक्षता 81.9% मापी गयी है[16]।

• Deep Learning Based Methodologies

विगत एक से डेढ़ दशक में न्यूरल नेटवर्क के विकास से Deep Learning के विभिन्न क्षेत्रों में कई महत्वपूर्ण अनुप्रयोग

सामने आये हैं। मशीन आधारित भाषा अनुवाद के क्षेत्र में भी Deep Learning का उपयोग हुआ है। Wisemano व अन्य ने Cluster based feature analysis का उपयोग करते हुए, Recurrent Neural Network पर आधारित Co-reference Resolution का एक नया मॉडल प्रस्तुत किया है[17]। इसी तरह Co-reference Resolution का अभी तक का सबसे अधिक दक्ष एवं बिना किसी Sytactic Parser के कार्य करने वाला, एक अन्य मॉडल ली एवं अन्य ने विकसित करने का दावा किया है[18]।

निष्कर्ष

उपरोक्त आलेख में मशीन आधारित भाषा अनुवाद से सम्बंधित एक जटिल समस्या ''संदर्भ निराकरण'' के बारे में विवरण दिया गया है। संदर्भ निराकरण की समस्या, अनुवाद में इसकी भूमिका, इसके प्रकार व सैद्धान्तिक समाधान के साथ सांख्यिकी आधारित कुछ महत्वपूर्ण एल्गोरिद्म्स के बारे में चर्चा की गयी है।

संदर्भ

1. Hirst Graham, "Anaphora in Natural Language Understanding", *Springer-Verlag*, Berlin, 1981.
2. MitkovRuslan, Choi Sung-Kwon &Sharp Randall, "Anaphora Resolution in Machine Translation", Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, Belgium, pp. 87-95, 1995.
3. Mitkov Ruslan, "Anaphora Resolution: The State of the Art", COLING'98/ACL'98 Tutorial on Anaphora Resolution; University of Wolverhampton, 1998.
4. Novak Marina, "Utilization of Anaphora in Machine Translation", WDS'11 Proceedings of Contributed Papers, 155-160, 2011.
5. Halliday Michael A. & HasanRuqaiya, "Cohesion in English", Longman English Language Series **9**, London, 1996.
6. Mitkov Ruslan, "Introduction: Special Issue on Anaphora Resolution in Machine Translation and Multilingual NLP", *Springer Journal on Machine Translation*,Volume **14**, Issue 3-4, pp. 159-161, 1999.
7. Seddik Khadiga Mahmoud & Ali Farghaly, "Anaphoric Relations", Natural Language Processing of Semitic Languages, 2014.
8. Shekhar Shivangi & Kumar Umesh, "Review on the Techniques of Anaphora Resolution", *International Journal of Latest Trends in Engineering and Technology*, Vol. **(8)** Issue (2), pp.037-040, 2018.
9. Sukthanker Rhea, Poria Soujanya, Cambria Erik & Thirunavukarasu Ramkumar, "A Review: Anaphora and Coreference Resolution" *arXiv* : 1805.11824, 2018.
10. Dagan Ido & Itai Alon, "Automatic Processing of Large Corpora for the Resolution of Anaphora References", Proceedings of the 13th International Conference on Computational Linguistics Vol. **III**, 1-3, Helsinki, Finland, 1990.
11. Nasukawa Tetsuya, "Robust Method of Pronoun Resolution Using Full-Text Information", Proceedings of the 15th International Conference on Computational Linguistics, (COLING'94), 1157-1163, Kyoto, Japan, 1994.
12. Mitkov Ruslan, "An Uncertainty Reasoning Approach for Anaphora Resolution", Proceedings of the Natural Language Processing Pacific Rim Symposium, pp. 149-154, Seoul, Korea,1995.
13. Hardmeier Christian & Federico Marcello, "Modelling Pronominal Anaphora in Statistical Machine Translation", In International Workshop on Spoken Language Translation; Paris, France, pp 283-289, December 2nd & 3rd, 2010.
14. Broscheit S, Poesio M, Ponzetto S. P., Rodriguez K, Joseba, Romano L O, Uryupina Y, Versley &Zanoli R, "BART: A Multilingual Anaphora Resolution System", In Proceedings of the 5th International Workshop On Semantic Evaluations (Sem Eval-2010), Uppsala, Sweden, 15-16 July 2010.
15. Kamune Kalyani P & Agrawal Avinash "Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text", *International Journal of Intelligent Systems and Applications*, **(02)** 56-64, 2015.
16. Charniak Eugene, "A Maximum Entropy Inspired Parser", In Proceedings of NAACL-2000, 1st North American chapter of the Association for Computational Linguistics conference, Pages 132-139, Seattle, Washington-April 29 - May 04, 2000.
17. Wiseman Sam, Rush Alexander M. & Shieber Stuart M, "Learning Global Features for Co & reference Resolution", *arXiv*:1604.03035, 2016.
18. Lee Kenton, He Luheng, Lewis Mike & Zettlemoyer Luke", End-to-end Neural Coreference Resolution", *arXiv* preprint *arXiv*: 1707.07045, 2017.