Journal of Advances and Scholarly Researches in Allied Education
Vol. XIV, Issue No. 1, October-2017, ISSN 2230-7540

हिंदी रूप विश्लेषक में शब्द-भेद टैगर की भूमिका

Dr. Gita Sahay*

Department of Hindi, Gaya College, Gaya

सारांश – ***शब्द-भेद टैगिंग (Part of Speech Tagging) प्राकृतिक भाषा संसाधन (Natural Language Processing) का एक महत्वपूर्ण भाग है। हिंदी भाषा के वैश्विक प्रयोग को देखते हुए आज भारतीय संदर्भ में मशीनी अनुवाद की गुणवत्ता में सुधार के लिए निरंतर शोध हो रहे है। रूप विश्लेषक (Morph Analyzer) मशीनी अनुवाद प्रक्रिया का वह हिस्सा है जिसके माध्यम से किसी भाषा में प्रयुक्त होने वाले शब्दों का सूक्ष्म विश्लेषण किया जाता है। इस विश्लेषण में सबसे पहले किसी शब्द की व्याकरणिक कोटि को ज्ञात किया जाता है। शब्द की व्याकरणिक कोटि को ज्ञात करने की इस प्रक्रिया को शब्द-भेद टैगिंग के नाम से जाना जाता है। हिंदी भाषा में संज्ञा, सर्वनाम, क्रिया, क्रिया विशेषण आदि शब्द-भेद की कोटियाँ है।***

मुख्य शब्दः रूप विश्लेषक, प्राकृतिक भाषा संसाधन, मशीनी अनुवाद, हिंदी, शब्द-भेद टैगिंग

----------X----------

परिचयः

वैश्विक परिप्रेक्ष्य में भाषा की महत्ता को देखते हुए द्वितीय विश्व युद्ध के दौरान मशीनी अनुवाद की संकल्पना अस्तित्व में आई। आज मशीनी अनुवाद इस स्थिति में पहुंच गया है कि हम इंटरनेट के माध्यम से एक भाषा की पाठ्य सामग्री को किसी दूसरी भाषा में मात्र कुछ सेकंडों में प्राप्त कर सकते हैं। इन मात्र कुछ सेकंडों में मशीन (कंप्यूटर) एक जटिल प्रक्रिया से गुजरती है। इस प्रक्रिया के दौरान वह दोनों भाषाओं के व्याकरणिक, संरचनात्मक, व्यवहारिक आदि सभी पक्षों का अध्ययन करती है और उसके बाद हमारे सामने एक अनुवाद प्रस्तुत होता है। प्राकृतिक भाषा संसाधन संगणकीय भाषा विज्ञान का एक क्षेत्र है जिसमें प्राकृतिक भाषाओं यथा हिंदी, अंग्रेजी, चाइनीज आदि को विश्लेषित किया जाता है तथा

उन्हें मशीन पठनीय रूप में संसाधित किया जाता है तथा इसके माध्यम से भिन्न-भिन्न भाषाई उपकरणों का निर्माण किया जाता है।

रूप विश्लेषक मशीनी अनुवाद प्रक्रिया का एक मुख्य उपकरण है। रूप विश्लेषण किसी भाषा में प्रयुक्त होने वाले शब्दों का भाषिक अध्ययन होता है जिसके द्वारा किसी शब्द की छोटी-से-छोटी भाषिक जानकारी प्राप्त की जाती है। इस भाषिक अध्ययन में किसी शब्द का धातु रूप, शब्द-भेद कोटि, लिंग, वचन, कारक, पुरुष, काल, पक्ष, वृत्ति आदि सूचनाएँ मशीन को प्रदान की जाती है। इन प्रदान की गई सूचनाओं के आधार पर ही मशीनी अनुवाद की गुणवत्ता निर्भर होती है।

शब्द-भेद के अंतर्गत किसी शब्द की व्याकरणिक कोटियों का अध्ययन किया जाता है। इसमें प्रत्येक शब्द की एक व्याकरणिक कोटि निर्धारित कर दी जाती है, जिसे उस शब्द का टैग कहा जाता है। शब्द-भेद के लिए निर्मित किए गए उपकरण को शब्द-भेद टैगर कहा जाता है। भारतीय परिप्रेक्ष्य में मशीनी अनुवाद की गुणवत्ता में सुधार हेतु भिन्न-भिन्न भाषाओं के शब्द-भेद टैगरों का निर्माण किया गया है। प्रस्तुत शोध पत्र में यूनिवर्सल डिपेंडेंसी आधारित हिंदी भाषा के लिए निर्मित रूप विश्लेषक में शब्द-भेद टैगर निर्माण की विधि का विवरण प्रस्तुत किया गया है।

साहित्य पुनरवलोकनः

हिंदी तथा अन्य भारतीय भाषाओं के शब्द-भेद टैगर निर्माण के लिए बहुत-से शोध किए गए हैं जिनमें नियम उपागम आधारित, सांख्यिकीय उपागम आधारित तथा संकर उपागम आधारित पद्धतियों का प्रयोग किया है। इस क्षेत्र में किए गए कुछ शोध निम्नांकित हैं:

1. शुभांगी राठोड तथा अन्य ने अपने शोध में भारतीय क्षेत्रीय भाषाओं के लिए विभिन्न शब्द-भेद टैगिंग तकनीकों की चर्चा की है। इन्होंने अपने शोध में

www.ignited.in

नियम आधारित, सांख्यिकीय आधारित तथा संकर आधारित पद्धतियों का वर्णन किया है।

2. प्रवेश कुमार तथा अन्य ने संकर उपागम का प्रयोग करते हुए एक हिंदी शब्द-भेद टैगर का निर्माण किया है। यह उपकरण 500 वाक्यों के कॉर्पस पर विश्लेषित किया गया है।

3. नवनीत गर्ग तथा अन्य ने नियम आधारित हिंदी शब्द-भेद टैगर प्रस्तुत किया है। यदि डेटाबेस में कोई टैग उपलब्ध नहीं होता है तो शब्द को टैग करने के लिए विभिन्न नियमों का प्रयोग किया जाता है। इस उपकरण का 26,149 शब्दों पर विश्लेषण किया गया है तथा इसकी शुद्धता 87.55: है।

4. अभिजीत पॉल तथा अन्य ने एचएमएम उपागम का प्रयोग करते हुए साख्यिकीय पद्धति से नेपाली भाषा का शब्द-भेद टैगर प्रस्तुत किया है। इसमें 1,50,839 शब्दों का प्रयोग किया गया है।

शब्द-भेद टैगर निर्माण की प्रविधियाँ:

शब्द-भेद टैगर निर्माण के लिए मुख्य रूप से तीन पद्धतियाँ प्रचलित है।

1. **नियम आधारित पद्धतिः** इस पद्धति में टैग की अशुद्धियों को दूर करने के लिए नियम बनाए जाते है। ये नियम वाक्य में किसी शब्द के पहले तथा बाद में प्रयोग किए जाने वाले शब्दों को ध्यान में रखते हुए बनाए जाते है। नियम बनाते समय संदर्भगत सूचना का भी ध्यान रखा जाता है। इसके लिए भाषा के व्याकरण तथा संदर्भगत नियमों की जानकारी की आवश्यकता होती है। जैसे:

नियम 1

राधा (संज्ञा) ने (परसर्ग) गाना गाया।

यदि किसी शब्द का टैग परसर्ग है तो उससे पहले शब्द का टैग संज्ञा होने की सर्वाधिक संभावना होती है।

नियम 2

विवेक को काला (विशेषण) घोडा (संज्ञा) पसंद है।

यदि किसी शब्द का टैग विशेषण है तो उसके बाद वाले शब्द का टैग संज्ञा होने की सर्वाधिक संभावना होगी।

2. **सांख्यिकीय आधारित पद्धतिः** इस पद्धति में किसी शब्द के टैग का निर्धारण उस शब्द के लिए सर्वाधिक प्रयोग किए गए टैग के आधार पर किया जाता है। इस प्रक्रिया में टैगर की शुद्धता डेटा पर आधारित होती है। इस पद्धति से निर्मित टैगर में डेटा की संख्या जितनी अधिक होती है टैगर की शुद्धता की संभावना भी उतनी अधिक बढ़ती जाती है।

3. **संकर आधारित पद्धतिः** यह पद्धति नियम आधारित तथा सांख्यिकीय आधारित पद्धति का मिलाजुला रूप होती है। इस पद्धति से निर्मित टैगर में व्याकरणिक नियमों तथा सांख्यिकी दोनों का प्रयोग किया जाता है। जहां पर भाषाई नियम विफल हो जाते हैं वहाँ पर इस पद्धति में सांख्यिकीय आधार पर टैग का निर्धारण किया जाता है। शब्द-भेद टैगर की शुद्धता के लिए इस पद्धति का सर्वाधिक प्रयोग किया जाता।

हिंदी रूप विश्लेषक के लिए शब्द-भेद टैगर निर्माणः

रूप विश्लेषक निर्माण के लिए शब्द को तीन प्रक्रियाओं से गुजरना होता हैः

1. धातु रूप

2. शब्द-भेद टैगिंग

3. रूप विश्लेषण टैगिंग

प्रस्तुत शोध-पत्र में शब्द-भेद टैगिंग प्रक्रिया के विभिन्न चरणों को विस्तार से प्रस्तुत करने का प्रयास किया गया है।

टैगसेटः शब्द-भेद टैगर के निर्माण के लिए सबसे पहले व्याकरण आधारित एक टैगसेट का निर्धारण किया जाता है। टैगसेट निर्माण में इस बात का ध्यान रखा जाता है कि किसी भाषा में प्रयोग किए जा रहे प्रत्येक शब्द के लिए हमारे पास एक निश्चित टैग अवश्य उपलब्ध होना चाहिए। यहाँ हिंदी भाषा के लिए शब्द-भेद टैगर निर्माण के लिए यूनिवर्सल डिपेंडेंसी द्वारा प्रयोग किए टैगसेट का प्रयोग किया गया है। इस टैगसेट में हिंदी भाषा की व्याकरणिक कोटियों के अनुसार 16 टैग निर्धारित किए गए हैं। इन सभी टैगों को उदाहरण सहित निम्नलिखित तालिका के माध्यम से समझा जा सकता हैः

क्र.सं.	व्याकरणिक कोटि	टैग	उदाहरण
2	विशेषण	ADJ	काला, अफ्रीकी, भारतीय, पुराना
2	परसर्ग	ADP	ने, को, से, द्वारा, लिए
3	क्रिया–विशेषण	ADV	लगातार, आगे, पहले, बाद
4	सहायक क्रिया	AUX	था, है, होंगे, हूँ
5	समानाधिकरण योजक	CCONJ	तथा, और, व, पर
6	निर्धारक	DET	वह, यह, सभी, कम
7	विस्मयादिबोधक	INTJ	अरे, वाह, आह, ऐ
8	संज्ञा	NOUN	कहानी, घड़ी, लड़के, घोड़ा
9	संख्या	NUM	एक, दो, हजार, लाख
10	अव्यय	PART	ही, तो, भी, नहीं,
11	सर्वनाम	PRON	मैं, तू, आप, वह
12	व्यक्तिवाचक संज्ञा	PROPN	राम, सीता, भारत, मुंबई
13	विराम चिह्न	PUNCT	?,।,(,)
14	आश्रित योजक	SCONJ	कि, अगर, बल्कि, यानि
15	क्रिया	VERB	खा, आ, जा, कर
16	अतिरिक्त	X	रिजल्ट, स्टंप, इंटरनेट, बॉल

डेटा संकलन: हिंदी के शब्द-भेद टैगर के लिए कहानी, उपन्यासों से 5000 वाक्यों का संकलन किया गयज्ञं इन सभी वाक्यों का संकलन हिंदी समय-कॉम वेबसाइट से किया गया है।

पद्धति चयनः हिंदी शब्द-भेद टैगर के निर्माण के लिए हिडन मार्कोव मॉडल (Hidden Markov Model, & HMM) का प्रयोग किया गया। एचएमएम एक सांख्यिकीय उपागम है। यह एक संभावना आधारित मॉडल है। इसमें टैग की पूर्व तथा पश्च संभावना की गणना की जाती है तथा किसी शब्द के लिए सबसे अधिक संभावित टैग दिया जाता है। टैग निर्धारण के लिए निम्नलिखित सूत्र का प्रयोग किया जाता है:

P (ti/wi)=P (ti/ti-1)]P (ti+1/ti)-P (wi/ti)

P (ti/ti-1) दिए गए पिछले टैग की वर्तमान टैग होने की संभावना है।

P (ti+1/ti) दिए गए वर्तमान टैग की अगला टैग होने की संभावना है।

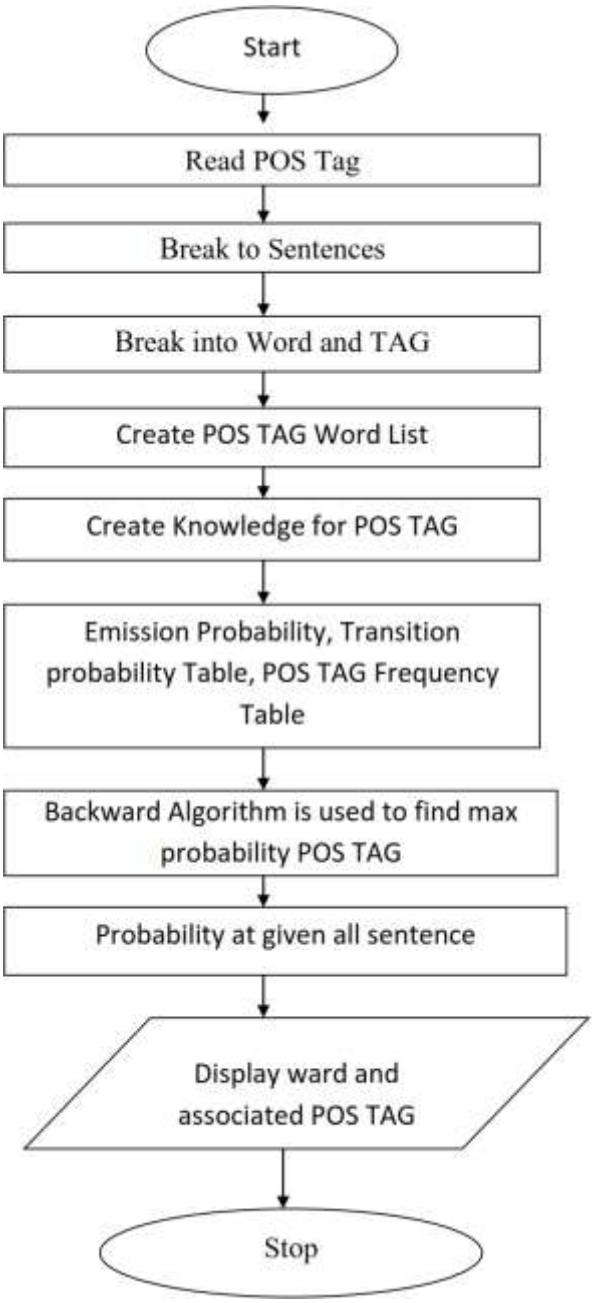
P (wi/ti) शब्द की वर्तमान टैग होने की संभावना है।

इन संभावनाओं की गणना निम्न सूत्र से की जाती हैः

P (ti/ti-1) = freq (ti&1, ti) / freq (ti&1)

प्रक्रियाः शब्द-भेद टैगर के निर्माण के लिए संकलित किए गए 5000 वाक्यों को दिए गए टैगसेट से टैग किया गया। इसके उपरांत एचएमएम मॉडल का प्रयोग करते हुए एक स्वचालित हिंदी शब्द-भेद टैगर का निर्माण किया गया। एचएमएम मॉडल उन 5000 वाक्यों किसी शब्द को सर्वाधिक बार जो टैग दिया गया है तथा उस शब्द से पहले प्रयोग किए गए शब्द के टैग तथा उस शब्द के बाद वाले शब्द के लिए प्रयोग किए गए टैग के आधार पर सांख्यिकीय गणना करता है और स्वचालित टैगर में इनपुट के रूप में दिए गए वाक्य के प्रत्येक शब्द को टैग करने के लिए सर्वाधिक उपयुक्त टैग का निर्धारण करता है।

सांख्यिकीय आधारित हिंदी शब्द-भेद टैगर के निर्माण के लिए ब्रु प्रोग्रामिंग भाषा का प्रयोग किया गया है तथा यह .दमज फ्रेमवर्क 4.5 पर कार्य करता है। इसकी निर्माण प्रक्रिया निम्नलिखित फ्लो चार्ट के माध्यम से समझा जा सकता हैः



परिणामः शब्द-भेद टैगर के निर्माण की प्रक्रिया पूर्ण होने के उपरांत टैगर में इनपुट के रूप में किसी हिंदी वाक्य को टैग करने के लिए दिया जाता है तो उक्त वाक्य का निम्नलिखित रूप में टैग आउटपुट प्राप्त होता हैः

इनपुट वाक्यः राम के घर में एक 200 साल पुरानी अलमारी रखी है।

टैग आउटपुटः राम\ PROPN के\ ADP घर\ NOUN में\ ADP एक\ NUM साल\ NOUN पुरानी\ ADJ अलमारी\ NOUN राखी\ VERB है\ AUXI/PUNCT

भविष्य में संभावनाएं: प्रस्तुत शोध पत्र में साहित्य के क्षेत्र से मात्र 5000 वाक्यों के आधार पर हिंदी शब्द-भेद टैगर का निर्माण किया गया है। भविष्य में इसी शोध को आगे बढ़ाते हुए भिन्न-भिन्न क्षेत्रों से 50000 वाक्यों का संकलन करके एक हिंदी रूप विश्लेषक के निर्माण की योजना है। वाक्यों की संख्या बढ़ने से शब्द-भेद टैगर की शुद्धता में भी वृद्धि होगी और मशीनी अनुवाद के लिए हिंदी रूप विश्लेषक निर्माण में सहायता मिलेगी।

संदर्भः

1. Jain, S- & Mishra N (2017). Insight of Various POS Tagging Techniques for Hindi Language, IJCSEITR, Volume 07, Pages 29-34.

2. Garg, N., Goyal, V. & Preet, S. (2012). Rule Based Part of Speech Tagger, Proceedings of COLING, Pages 163-174.

3. Joshi, N, Darbari, H- & Mathur, L. (2013). HMM Based POS Tagger for Hindi, Computer Science & Information Technology, Pages 341-349.

4. Goyal, V. & Lehal, G.S. (2008). Hindi Morphological Analyser and Generator- Proceedings of the First International Conference on Emerging Trends in Engineering and Technology, Pages 1156-1159.

5. Rastogi, M. & Khanna, P. (2014). Development of Morphological Analyser for Hindi. International Journal of Computer Applications, Volume 95.

6. Singh,P.D., Pramila, Singh, SP., Kumar A. & Hemant Darbari, (March 2014). Morphological Analyser for Hindi A Rule Based Implementation- International Journal of Advanced Computer Research.

7. Dwivedi, P.K. & Malakar, P.K. (2015). Hybrid Approach Based POS Tagger for Hindi Language. IJRSCSE, Pages 63-68.

Corresponding Author

Dr. Gita Sahay*

Department of Hindi, Gaya College, Gaya