Relational Learning Assisted Construction of Rule Base for Indian Language NER

Anup Patel **Ganesh Ramakrishnan** **Pushpak Bhattacharya**

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay, India
{anuppatel,ganesh,pb}@cse.iitb.ac.in

Abstract

We present Inductive Logic Programming (ILP) based techniques for automatically extracting rules for Named Entity Recognition (NER) from tagged corpora and background knowledge. Results using WARMR (Luc Dehaspe and Luc De Raedt 1997) and TILDE (Hendrik Blockeel and Luc De Raedt 1998) to learn rules for named entities of Hindi[1] and Marathi[2] show that the ILP approach has two advantages over hand-crafting the NER rules: (i) the *development time* reduces by a factor of 120 compared to a linguist doing the entire rule development, and (ii) a *complete and consistent view* of all significant patterns in the data at the level of abstraction specified through the mode declarations prevails in the learned rules.

1 Introduction

Named entity recognition- a critical NLP task- was first introduced in the sixth Message Understanding Competition (R Grishman and B Sundheim 1996) and consisted of three detection subtasks:

a) Proper names and acronyms of persons, locations, and organizations (ENAMEX)
b) Absolute temporal terms (TIMEX)
c) Monetary and other numeric expressions (NUMEX).

Early named entity recognition systems were rule-based with hand-crafted rules (D E Appelt, et al. 1993). Since hand-crafting of rules is tedious, algorithms for automatic learning rules were developed (M E Califf and R J Mooney 1999, S Soderland 1999), but these approaches did not provide adequate mechanisms for incorporating linguistic knowledge.

In this paper we show an Inductive Logic Programming based mechanism for NER rule extraction from NE tagged corpora. Our motivation has been to give computational support to a linguist in her task of formulating the NER rules.

This paper is organized as follows: Section 2 describes the complexity of Named Entity Recognition for Indian Languages, the motivation for using an ILP approach for this task and some specifics of the ILP approach. In Section 3, we describe our way of representing named entity tagged data in first order logic. In Section 4 we present our experimental results for the ILP and other approaches on Indian Language NER. In Section 5 we show our analysis of rules given by the ILP approach. Finally in Section 6 we conclude and propose future work in this direction.

2 NER for Indian Languages using ILP

For Indian languages we don't have the privilege of huge tagged corpus. Table 1 below shows the current status of tagged corpus for NER in Hindi and Marathi.

Table 1: Hindi and Marathi named entity corpus

	Marathi	**Hindi**
Sentences	3884	22748
Words	54340	547138
Person Tags	3025	5253
Organization Tags	833	2473
Location Tags	997	6041

Compounded with the limitations of the paucity of tagged corpora, is the challenge of inherent ambiguity of NER task. Table 2 illustrates some of these ambiguities using Marathi as the example language.

[1] Hindi is the official national language of India. The language and its close cousin Urdu are spoken by approximately 500 million people in the world.

[2] Marathi is the official language of Maharashtra, a state in Western India. The language has close to 20 million speakers in the world.

Proceedings of ICON 2009: 7th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009

Inductive Logic Programming (ILP) (S. H. Muggleton 1991), deals with learning from instances of objects represented in a relational form. Several authors have used ILP, or ILP-inspired systems for information extraction. Notable amongst these are: Use of ILP to construct theories for IE (J. S. Aitken 2002); Califf's work with Rapier (M E Califf and R J Mooney 1999), which is inspired by bottom-up ILP systems; and the work of Roth and colleagues (Dan Roth and Wen tau Yih 2001) who use restricted templates defined by "relation generating functions" to construct features for IE. Our results here are intended to add these by providing evidence for the following: *general-purpose ILP systems can enable efficient construction of a consistent rule-based system for Indian language named entity recognition.*

There are number of ways in which we can use the rules learned by ILP, but for simplicity we show three ways of consolidating learned rules:

a) Retain the default ordering of learned rules in the rule firing engine.

b) Induce an ordering on the learned rules using greedy heuristics such as in (Venkatesan Chakravarthy, et al. 2008).

c) Construct a feature corresponding to each rule, with the feature value 1 if the rule covers an instance and 0 otherwise. The features (which can be functions of both the head as well as the body of the rules) can be used in a statistical graphical model such as CRF (John Lafferty, Andre McCallum and Fernando Pereira 2001). The need for a graphical model is driven by our need for structured learning:

   a. The named entity disambiguation of a token can be potentially influenced by the entity disambiguation of adjacent tokens.

   b. The features (obtained as transformation of the rules), are functions of the input (token sequence) as well as of the output (possible labels that can be associated with the current and adjacent tokens). Models such as support vector machines and Naïve Bayes classifiers can only handle features that are functions of the input.

We have experimented with two ILP techniques:

1. **WARMR:** This is an extension of the apriori algorithm to first-order logic. Typically apriori based techniques are computationally expensive and the resulting rules are not ordered. We need to explicitly induce ordering using some heuristic or greedy approach. We use consolidation techniques **b)** and **c)** in this case because ordering a set of rules is a NP-hard problem (Venkatesan Chakravarthy, et al. 2008).
2. **TILDE:** This is an extension of traditional C4.5 decision tree learner to first-order logic. Decision tree induction algorithms are usually greedy and hence computationally faster than WARMR like algorithms. We use consolidation technique **a)** in this case because the set of rules (decision list) output by TILDE are already ordered.

Table 2: Ambiguities in named entities found in Marathi

Ambiguity	**Examples**
Variations of Proper Nouns	• डॉ. काशिनाथ घाणेकर, डॉ. घाणेकर *(Dr. Kashinath Ghanekar, Dr. Ghanekar)* • भारतीय जनता पार्टी, भा. ज. पा. *(Bhartiya Janta Party, B. J. P.)*
Person v/s Adjective v/s Verb	• डॉ. लागू/PER यांनी मनोगत मांडले *(Dr. Lagu expressed his thoughts)* • ही योजना संपूर्ण शहरात लागू/JJ करण्यात येणार आहे. *(This scheme will be applicable in the whole city.)* • ..... पण अजिबात झोप लागू/VM दिली नाही. *(..... but he didn't allow me fall asleep at all.)*
Person v/s Common Noun	• मुंबईला आल्यावर डॉक्टरांना/PER फोन करणे भागच होते. *(After coming to Mumbai it was must to call the Doctor.)* • तू डॉक्टर/NN की मी? *(Are you doctor or me?)*
Person v/s Organization	• नेताजींच्या/PER गूढ मृत्यूचा मागोवा ..... *(Following Netaji's suspicious death .....)*

	• "मिशन नेताजी/ORG' या स्वयंसेवी संस्थेने ..... *(“Mission Netaji” is a voluntary organization that .....)*
Person v/s Facility	• सरस्वती आणि लक्ष्मीची/PER एकत्रित उपासना केल्यास ..... *(If Saraswati and Laxmi are worshiped together .....)* • श्रीकृष्ण,सुंदर, लक्ष्मी/FACअशी नाट्य मंदिरे होती. *(There were Drama Theaters like Shri Krishna, Sundar, Laxmi.)*
Organization v/s Common Noun	• ..... विनोद गपाटयांनी "सकाळ'शी/ORG बोलताना सांगितले *(Vinod Gapte while talking with Sakal (newspaper) said .....)* • सकाळपासून/NN त्यांना शुभेच्छा देणारे दूरध्वनी खणखणत होते . *(Many calls are coming from morning to congratulate him.)*
Organization v/s Location	• पाक/ORG संघ/ORG शनिवारी लंडनमार्गे पाकला/LOC प्रयाण करणार आहे. *(Pakistan team will go to Pakistan via London on Saturday)*
Location v/s Person	• निगडी येथील भक्ती शक्ती चौक, टिळक/LOC चौक/LOC, ... *(Bhakti Chauk, Tilak Chauk, ... from Nigdi)* • टिळक/PER व डॉ. बाबासाहेब आंबेडकर ..... *(Tilak and Dr. Ambedkar .....)*
Location v/s Date	• बुधवार पेठ/LOC येथील आर. जी. कंपनीने ..... *(R.J. Company from Budhavar Peth .....)* • या समितीचे काम बुधवारपासून/DAT सुरू झाले. *(Committee's work has started from Wednesday.)*

(**Note:** ORG=Organization, PER=Person, FAC=Facility, LOC=Location, DAT=Date, NN=Noun, JJ=Adjective, and VM=Verb. In above examples ambiguous entities are shown in red.)

3 Representing named entity data in first order logic

Most ILP systems require input examples, background knowledge and mode declarations in the form of first order predicates. Therefore, to learn rules for NER we first convert tagged data into first order logic. We create first order logic data from Hindi and Marathi tagged data as follow:

i. **Input Examples:** We will have one input example for each word of each sentence from the corpus. Each input example is a set of predicates describing a set of properties of the word and surrounding words in a window of size one. Each example will have unique identifier and properties of words are represented by 3-ary predicates. The first argument of each predicate is the unique identifier for example, second argument is relative position of word whose property we are describing and third argument is value of the property. As an illustration, consider the input example shown in Figure 1 (d) for word काशिनाथ in the sample Marathi sentence shown in Figure 1 (a). For simplicity we have shown only four predicates describing properties of words, but in our implementation we have used many more predicates.

ii. **Background Knowledge:** In background knowledge we assert more facts about the constants appearing as third argument of the predicates used in input examples. For simplicity we have used only unary predicates in our representation but in general any horn clause can be used. Figure 1 (b) shows a sample background knowledge created for the sample sentence shown in Figure 1 (a).

iii. **Mode declarations:** In most ILP systems mode declarations are represented using built-in predicates, which vary from system to system. These mode declarations restrict hypothesis search space for ILP systems and also control the predicates appearing in the learned rules. In our case predicate *p_entity(X,0,...)* should appear in the head of learned rule and other predicates in the body of learned rule. Figure 1(c) shows example of a learned rule.

डॉ. काशिनाथ घाणेकर भारावल्यासारखे माझ्या आईसमोर नमस्कारासाठी वाकले.
Getting carried away, Dr. Kashinath Ghanekar bowed in front of my mother.

(a) Sample Marathi Sentence

b_entity(PER). ***b_entity***(ORG). ***b_entity***(LOC). ...
b_word(डॉ.). ***b_word***(काशिनाथ). ***b_word***(घाणेकर). ...
b_postag(NNP). ***b_postag***(NN). ***b_postag***(DT). ...
b_wordcollections(titles). ***b_wordcollections***(firstnames). ...

(b) Sample Background Knowledge

p_entity(X,0,PER) :- ***p_word***(X,-1,डॉ.), ***p_postag***(X,0,NNP).

(c) Example of learned rule based on mode declarations

p_entity(d0s10w1,-1, PER).
p_postag(d0s10w1,-1,NNP).
p_word(d0s10w1,-1, डॉ.).
p_wordcollections(d0s10w1,-1,titles).
p_entity(d0s10w1,0, PER).
p_postag(d0s10w1,0, NNP).
p_word(d0s10w1,0, काशिनाथ).
p_wordcollections(d0s10w1,0, firstnames).
p_entity(d0s10w1,1, PER).
p_postag(d0s10w1,1,NNP)
p_word(d0s10w1,1, घाणेकर).
p_wordcollections(d0s10w1,1, lastnames).

(d) Input Example for word काशिनाथ

Figure 1: An input example for word in the sample sentence

4 Experimental Results

We have used a hand-crafted rule based named-entity recognizer for Marathi and Hindi developed by a linguist using the GATE (Hamish Cunningham, et al. 2002) system. The rules were hand-crafted over a period of 1 month (240 hours for 8 hours per day). We measured the performance of hand-crafted rule based system on 20% of tagged corpus for both Hindi and Marathi. This hand-crafted rule based systems will be our baseline system for comparison.

Parallelly, we learnt Marathi and Hindi named entity rules using the WARMR and TILDE systems available as a part of ACE (Hendrik Blockeel, ACE Datamining System 2008) data mining system over 80% of tagged corpus. For both systems we used a common minimum support threshold of 20 examples (for Marathi) and 50 examples (for Hindi). As explained before each example for our experiments contains all words and their properties in window of size one. Unfortunately due to lack of sufficient computational resources we were not able to use WARMR system for rule induction over Hindi. The Table 3 below summarizes time taken by rule induction process.

To compare quality of the learnt rules we consolidated and apply them over the remaining 20% of the tagged corpus in following ways:

1. **TILDE Rule Based NER:** Rules learned by TILDE are plugged in a rule-based named entity recognizer without altering the order of rules.
2. **WARMR Rule Based NER:** Rules learned by WARMR are ordered using simple precision score heuristic and a greedy algorithm mentioned in (Venkatesan Chakravarthy, et al. 2008). These ordered rules are then plugged into a rule-based named entity recognizer.
3. **WARMR CRF Based NER:** Rules learned by WARMR plugged into CRF (Sunita Sarawagi 2004) as features ignoring the order of rules.

The performances of the hand-crafted rule based (HR), the TILDE rule based (TR), the WARMR rule based (WR), and the WARMR CRF based (WC) systems are shown below in Table 4 (for Hindi) and Table 5 (for Marathi).

Table 3: Time taken for Rule Induction Process

Rule Induction Method	Time Taken (Hours)		Speed-Up (w.r.t. Hand-Craft)	
	Marathi	Hindi	Marathi	Hindi
Hand-Craft	240	300	1	1
WARMR	140	--	1.7	--
TILDE	2	4	120	75

Table 4: Experimental results for Hindi

Entity	Precision		Recall		F-Measure	
	HR	**TR**	**HR**	**TR**	**HR**	**TR**
PER	0.63	**0.73**	0.39	**0.62**	0.48	**0.67**
ORG	0.69	**0.72**	0.11	**0.42**	0.19	**0.53**
LOC	0.60	**0.82**	0.56	**0.62**	0.58	**0.71**

Table 5: Experimental results for Marathi

Entity	Precision				Recall				F-Measure			
	HR	**TR**	**WR**	**WC**	**HR**	**TR**	**WR**	**WC**	**HR**	**TR**	**WR**	**WC**
PER	0.61	0.55	0.60	**0.74**	0.70	**0.99**	0.90	0.91	0.65	0.71	0.72	**0.82**
ORG	0.15	**0.85**	0.19	0.59	0.10	0.37	0.46	**0.52**	0.12	0.51	0.27	**0.55**
LOC	0.51	**0.54**	0.41	0.51	0.24	0.18	0.35	**0.45**	0.33	0.27	0.38	**0.48**

5 Diagnosis of Induced Rules

A detailed comparison of hand-crafted rules and induced rules has shown that many of the hand-crafted rules were also discovered by the ILP rule induction process (shown in Table 6).

Table 6: Good induced rules

ID	Rule
M1	IF (Previous word has a dot ".") AND (Next word is यांनी [3]) THEN (Current word has PER tag)

M2	IF (Previous word is भारतीय [4])
	THEN (Current word has ORG tag)

H1	IF (Previous word is श्री [5])
	THEN (Current word has PER tag)

H2	IF (Current word has POS tag NNP)
	AND (Current word is a known location)
	AND (Next word is में [6])
	THEN (Current word has LOC tag)

Table 8: Over generalization/specialization

ID	Rule

M3	IF (Current word has POS tag NNP)
	THEN (Current word has ORG tag)

M4	IF (Current word has POS tag NNP)
	AND (Current word is पुणे)
	THEN (Current word has LOC tag)

At the same time in tables 7 to 9, we have reported problems in the induced rules arising from *limitations of data, hypothesis* and *background knowledge* and also over generalization and over specification.

Table 7: Hypothesis language limitation

ID	Rule
H4	IF {(Previous word is a first name) OR (Previous word is a last name)} AND (Current word has POS tag NNP) AND {(Next word is का) OR (Next word is ने)} THEN (Current word has PER tag)
H5	IF (Previous word is a first name) AND (Current word has POS tag NNP) AND (Next word is का) THEN (Current word has PER tag)
H6	IF (Previous word is a last name) AND (Current word has POS tag NNP) AND (Next word is का) THEN (Current word has PER tag)
H7	IF (Previous word is a first name) AND (Current word has POS tag NNP) AND (Next word is ने) THEN (Current word has PER tag)
H8	IF (Previous word is a last name) AND (Current word has POS tag NNP) AND (Next word is ने) THEN (Current word has PER tag)

[3] यांनी is a demonstrative pronoun [used in Marathi]

[4] भारतीय = *Indian* [used in both Hindi and Marathi]

[5] श्री = *Mr.* [used as person title in Hindi]

[6] में = *in* [used as postposition in Hindi]

Table 9: Data problem

ID	Rule

H3	IF (Current word has POS tag NNP)
	AND (Previous word is पूर्वी)
	THEN (Current word has LOC tag)

Table 10: Background knowledge limitation

ID	Rule
M5	IF (Current word paradigm is unknown) AND (Current word suffix is empty) AND (Next word has POS tag VM) THEN (Current word has LOC tag)

6 Conclusions

We have reported our work on creating NER systems for Hindi and Marathi, inducing rules in the ILP framework from annotated corpora. We note that the system which feeds Warmer-induced rules to a CRF system performs the best in the sense of highest F-score. This is not very surprising. CRF is a powerful probabilistic reasoning system; augmented with features as powerful as horn clauses, they can act as high accuracy sequence labelers. As mentioned already the same experiment for Hindi could not be preformed due to resource limitations.

Our future work consists of finding efficient rule induction methods on large volumes of annotated data, developing interactive ILP assisted rule development system for linguists, and include other languages.

Reference

D E Appelt, J R Hobbs, J Bear, D J Israel, and M Tyson. "Fastus: A finite-state processor for information extraction from real-world text." *IJCAI.* 1993. 1172–1178.

Dan Roth, and Wen tau Yih. "Relational learning via propositional algorithms: An information extraction case study." *In IJCAI.* 2001. pages 1257-1263.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. "Gate: An architecture for development of robust HLT applications." *Recent Advanced in Language Processing.* 2002. 168-175.

Hendrik Blockeel. "ACE Datamining System." *Machine Learning Group.* March 2008. http://www.cs.kuleuven.be/~dtai/ACE/doc/ACEuser-1.2.12-r1.pdf.

Hendrik Blockeel, and Luc De Raedt. "Top-down induction of logical decision trees." *Artificial Intelligence.* 1998.

J. S. Aitken. "Learning Information Extraction Rules: An Inductive Logic Programming approach." *In Proceedings of the 15th European Conference on Artificial Intelligence.* 2002. pages 355-359.

John Lafferty, Andre McCallum, and Fernando Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." *Proceedings of the International Conference on Machine Learning (ICML-2001).* 2001.

Luc Dehaspe, and Luc De Raedt. "Mining association rules in multiple relations." *Proceedings of the 7th International Workshop on Inductive Logic Programming.* 1997. 125-132.

M E Califf, and R J Mooney. "Relational learning of pattern-match rules for information extraction." *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99).* 1999. 328–334.

R Grishman, and B Sundheim. "Message understanding conference-6: A brief history." 1996. 466–471.

S Soderland. "Learning information extraction rules for semi-structured and free text." *Machine Learning.* 1999.

S. H. Muggleton. "Inductive Logic Programming." *New Generation Computing.* 1991. 8(4):295-318.

Sunita Sarawagi. *CRF Project Page.* 2004. http://crf.sourceforge.net/.

Venkatesan Chakravarthy, Sachindra Joshi, Ganesh Ramakrishnan, Shantanu Godbole, and Sreeram Balakrishnan. "Learning Decision Lists with Known Rules for Text Mining." *The Third International Joint Conference on Natural Language Processing (IJCNLP 2008).* 2008.