Resolving Actor Coreferences in Hindi Narrative Text

Nitin Ramrakhiyani Swapnil Hingmire Sachin Pawar

Sangameshwar Patil Girish K. Palshikar
{nitin.ramrakhiyani,swapnil.hingmire,sachin7.p}@tcs.com
{sangameshwar.patil,gk.palshikar}@tcs.com
TCS Research, Tata Consultancy Services, India

Pushpak Bhattacharyya
pb@cse.iitb.ac.in
IIT Patna, India

Vasudeva Varma
vv@iiit.ac.in
IIIT Hyderabad, India

Abstract

An important aspect of understanding narrative text is identification of actors, its mentions and coreferences among them. Coreference Resolution in Hindi is a relatively under-explored area. In this paper, we focus on the task of resolving coreferences of actor mentions in Hindi narrative text. We propose a linguistically grounded approach for the task using Markov Logic Networks (MLN). Our approach outperforms two strong baselines on a publicly available dataset and 4 other manually created datasets.

[सरदार पटेल]$_{A_1}$ का जन्म [गुजरात]$_{A_2}$ में 1875 में हुआ था। [वे]$_{A_1}$ [झवेरभाई पटेल]$_{A_3}$ एवं [लाडबा देवी]$_{A_4}$ की [चौथी संतान]$_{A_1}$ थे। [लन्दन]$_{A_5}$ जाकर [पटेल]$_{A_1}$ ने बैरिस्टरी की पढाई की और वापस आकर [अहमदाबाद]$_{A_6}$ में वकालत करने लगे। [महात्मा गांधी]$_{A_7}$ के विचारों से प्रेरित होकर उन्होंने [भारत]$_{A_8}$ के स्वतन्त्रता आन्दोलन में भाग लिया। स्वतन्त्रता आन्दोलन में [सरदार पटेल]$_{A_1}$ का सबसे पहला और बड़ा योगदान [खेडा]$_{A_9}$ संघर्ष में हुआ। [किसानों]$_{A_{10}}$ ने [अंग्रेज सरकार]$_{A_{11}}$ से भारी कर में छूट की मांग की। जब यह स्वीकार नहीं किया गया तो [सरदार पटेल]$_{A_1}$, [गांधीजी]$_{A_7}$ एवं [अन्य लोगों]$_{A_{12}}$ ने [किसानों]$_{A_{10}}$ से मुलाकात की और [उन्हे]$_{A_{10}}$ कर न देने के लिये प्रेरित किया।

Table 1: Sample Hindi narrative. Actor mentions are marked with [. . .] and mentions of the i^{th} actor are denoted by the subscript A_i.

1 Introduction

Narrative text describes related sequences of events involving a set of actors and interactions among them. The first step towards understanding of narrative text is to identify the actors involved and to resolve their coreferences. We define an actor to be an entity of type PERSON, LOCATION, or ORGANIZATION. These actors are referred in text through their mentions which can be of three types: named entities, *generic noun phrases (NPs)*[1] and pronouns (Walker et al., 2006).

In this paper, we aim to resolve coreferences of actor mentions in Hindi narrative text. We assume availability of gold-standard actor mentions and their types; and focus only on resolving coreferences among actor mentions. Unlike much of the earlier work, we do not restrict the coreferences to only pronouns (Anaphora Resolution) and their nominal antecedents. In addition to pronouns, we also consider generic NPs for coreference resolution. For instance, referring to the sample narrative in Table 1, we want to identify that various mentions like the named entities (सरदार पटेल, पटेल), the pronouns (वे , उन्होंने) as well as the generic NP (चौथी संतान) all refer to the same actor (सरदार पटेल).

Coreference resolution is known to be a challenging NLP problem. Even for languages such as English, which have good quality linguistic resources and datasets, coreference resolution has proved to be a hard problem (Ng, 2017). For languages which are relatively resource-poor such as Hindi, the problem gets exacerbated.

Some of the approaches in coreference resolution in Hindi are adapted from coreference resolution approaches for English. For example, Dutta et al. (2008) adapt the well-known Hobb's algorithm for Hindi. Certain approaches involve application of linguistic knowledge for co-reference resolution. Agarwal et al. (2007) propose an approach based on matching constraints for the grammatical attributes of different words while Prasad and Strube (2000) and Uppalapu and Sharma (2009) apply centering theory (Grosz et al., 1995) for

[1] A generic NP is a noun phrase which has a common noun as its head-word.

Proc. of ICON-2018, Patiala, India. December 2018 ©2018 NLPAI, pages 50–58

coreference resolution. Dakwale (2014) proposes a hybrid approach based on dependency structure and linguistics constraints to resolve pronominal references. Mujadia (2017) proposes a hybrid approach based on Paninian dependency grammar, linguistic rules and resources like DBPedia and word-embeddings to resolve nominal coreferences.

Note that major focus of the work in coreference resolution for Hindi has been on anaphora resolution. Anaphora resolution is a subset of more general coreference resolution problem. Anaphora resolution focuses on connecting pronouns to their antecedents which refer to the same entity. It does not focus on connecting a generic NP to its antecedent(s). Further, prior work for coreference resolution in Hindi uses supervised learning algorithms which need labeled training data to induce the classifier(s).

To resolve coreferences among actor mentions in Hindi narratives, we develop unsupervised algorithms based on Markov Logic Networks (MLN) (Domingos and Lowd, 2009). MLNs combine first order logic rules with probabilistic graphical models in a single representation. We encode linguistic knowledge relevant to coreferences in an MLN and use the inference in the MLN for coreference resolution. Thus, the approach is unsupervised, avoiding the need for labeled training data.

Major contributions of this work are: i) To the best of our knowledge, this is the first attempt at actor coreference resolution for Hindi narrative text, ii) An unsupervised approach based on Markov Logic Networks for coreference resolution in Hindi, and iii) A set of robust linguistic rules encoded in MLN, despite the absence of good NLP pre-processing tools (e.g., no constituency parser or semantic role labeller for Hindi). The paper is organized as follows: Section 2 describes the related work, Section 3 covers the details of our MLN-based coreference resolution approach, Section 4 describes the experimental analysis and Section 5 concludes the work with some pointers on future work.

2 Related Work

Coreference resolution is an extensively studied problem in computational linguistics. Several authors have proposed methods for coreference resolution. These methods can be broadly classified into three types of approaches: i) rule based methods, ii) machine learning based methods and iii) hybrid methods.

Rule based methods like the Hobb's algorithm (Hobbs, 1986) represent linguistic knowledge about coreference in the form of rules which are then used for coreference resolution. These linguistically motivated rules try to model various factors of coreference resolution such as gender agreement, number agreement, semantic relations like IS-A, semantic similarity, proximity or theories like centering theory based choice of referring expression (Grosz et al., 1995). A key limitation of such rule based approaches is that they require extensive human efforts to represent and process linguistic knowledge.

Machine learning based methods on the other hand are “knowledge-poor” methods (See (Ng, 2017) for an overview). These methods use a labelled corpus to train models for coreference resolution. Recently, several authors have proposed neural methods of coreference resolution, e.g. (Clark and Manning, 2016; Lee et al., 2017). Though, neural methods have shown promising coreference resolution results as compared to other learning methods, they need a large amount of labelled data and computational resources. Hence, they can not be applied to low-resource Indian languages for which a large coreference annotated data is expensive to obtain.

In the context of coreference resolution in Hindi texts several authors adapted methods for coreference resolution in English. Dutta et al. (2008) adapted the Hobb's algorithm, while Prasad and Srube (2000) and Uppalapu and Sharma (2009) adapted centering theory based coreference resolution for Hindi. Dakwale (2014) proposes a hybrid approach which first applies a set of rules on syntactic information of sentences and then incorporates grammatical and semantic information into supervised learning methods to resolve more ambiguous instances. It is important to note that most of the work for coreference resolution in Hindi are focused on resolution of pronominal references and not on generic NPs discussed earlier. Recently, Mujadia (2017) proposes a sieve-based hybrid approach for coreference resolution of pronouns as well as nominal references. The approach uses a set of sieves using Paninian Dependency Grammar, POS labels, morphology and animacy features, linguistic resources like Hindi WordNet, DBpedia derived named dictionary, Word2Vec and GloVe based word embeddings. However, this approach is supervised and hence, needs a labelled dataset.

A recent work by Patil et al. (2018) is the closest to our proposed approach. They use an MLN-based approach for resolving actor mention coreferences in English narrative text. They build upon the output from Stanford CoreNLP coreference resolution, whereas we attempt to address the problem from scratch.

3 Coreference Resolution using MLN

We propose a linguistically motivated approach for resolving coreferences. As it is difficult and effort-intensive to develop annotated datasets for Hindi Coreference resolution, we develop an unsupervised approach using Markov Logic Networks (MLN). MLNs combine logic with probabilistic graphical models. MLN allows representation of linguistic knowledge characterizing coreferences, in the form of weighted first order logic rules. Weight associated with each rule represents its strength. An MLN is constructed for a narrative which encodes multiple pieces of information regarding actor mentions in the narrative. It also includes the first order logic rules encoding the linguistic knowledge. Inference in such an MLN leads to the most likely coreference links among the actor mentions, while ensuring maximum weighted satisfiability of the linguistic rules.

3.1 Predicates

We design several *predicates* which are needed to represent the linguistic knowledge in the form of first order logic rules. There are two types of predicates:

- **Evidence predicates**: These predicates encode observed information regarding actor mentions and their relationships in a given narrative. Truth values are known for all groundings of such predicates.

- **Query predicates**: Truth values for all or some of the groundings of these predicates are unknown. Inference in MLN is needed to know the most likely truth values for these predicates.

Table 2 describes in detail various predicates used in the MLN rules.

3.2 Linguistic Rules

We express linguistic knowledge characterizing coreferences in the form of first order logic rules in the MLN. In addition to these rules, *evidences* are provided to the MLN in the form of observed true groundings of all evidence predicates in a given narrative.

Ensuring Equivalence of Coreferences: The query predicate $Coref$ represents the coreference relations among actor mentions, which is required to be an equivalence relation. Hence, we include following 3 rules:

Reflexivity: $Coref(x, x)$.
Symmetry: $Coref(x, y) \Rightarrow Coref(y, x)$.
Transitivity: $Coref(x, y) \wedge Coref(y, x) \Rightarrow Coref(x, z)$.

Actor Type Consistency: A necessary condition for any two actor mentions to be coreferences of each other is that their Actor/Entity types should be same. E.g., संतान *(child)* and समिती *(committee)* can never be coreferences of each other because actor type of the former is PERSON whereas actor type of the later is ORGANIZATION. The rule is expressed in first order logic as:

$$NER(x, t) \wedge NER(y, w) \wedge (t \neq w) \Rightarrow \neg Coref(x, y)$$

Identical Actor Mentions: If $Identical(x, y)$ is true for any pair of actor mentions, then x and y are likely to be coreferences. This is a high confidence rule and hence is associated with infinite weight.

$$Identical(x, y) \Rightarrow Coref(x, y).$$

Actual pairs of such actor mentions are provided as evidences to MLN. E.g., In the sentence सरदार पटेल स्वतंत्रता सेनानी थे। *(Sardar Patel was a freedom fighter.)*[2], the actor mention स्वतंत्रता सेनानी *(freedom fighter)* is **predicative nominal** of the subject सरदार पटेल. Such actor mentions generally refer to the same real life actor and these are often connected through a copula verb. The evidence provided here is:

$$Identical(\text{सरदार पटेल}, \text{स्वतंत्रता सेनानी})$$

There are some other copula-like verbs (such as बने *(became)*) which connect coreferent actor mentions. E.g., सरदार पटेल उप-प्रधानमंत्री बने। *(Sardar Patel became the Deputy Prime Minister.)* Here, the evidence would be:

$$Identical(\text{सरदार पटेल}, \text{उप-प्रधानमंत्री})$$

[2] The English translations are provided only for reading help, the rules have to be interpreted for Hindi sentences only.

Evidence predicates:	
$NER(x,t)$	True iff actor mention x is of type t
$PronounLike(x)$	True iff actor mention x is a pronoun or a definite mention
$Identical(x,y)$	True iff actor mentions x and y appear in a linguistic relationship which indicates that x and y are likely to be coreferences of each other
$LexSim(x,y)$	True iff actor mentions x and y lexicall similar, i.e. they have little edit distance or one is suffix/prefix of another
$NonIdentical(x,y)$	True iff actor mentions x and y appear in a linguistic relationship which indicates that x and y are not likely to be coreferences of each other
$IsAntecedent(x,y)$	True iff actor mention y is an antecedent for x which is a pronoun or a definite mention
Query predicates:	
$Coref(x,y)$	True iff actor mentions x and y are coreferences of each other

Table 2: Predicates used in MLN rules

Lexically Similar Mentions: If x and y are lexically similar then they are likely to be coreferences. This rule is associated with a finite weight of 10 (empirically decided using a development dataset).

$$10.0\ LexSim(x,y) \Rightarrow Coref(x,y).$$

Actual pairs of such actor mentions are provided as evidences to the MLN. Following is an example of such an evidence:

$LexSim$(सरदार पटेल, सरदार वल्लभभाई पटेल)

Non-identical Actor Mentions: If $NonIdentical(x,y)$ is true for a pair of actor mentions (x,y), then x and y are **unlikely** to be coreferences. This is a high confidence rule and hence is associated with infinite weight.

$$NonIdentical(x,y) \Rightarrow \neg Coref(x,y).$$

Actual pairs of such actor mentions are provided as evidences to MLN. There are several cases for identifying non-identical pairs of actor mentions.

1. **Conjunctions**: When two actor mentions are conjunctions of each other, then it is highly unlikely that they are coreferences of each other. E.g., सरदार पटेल और अन्य नेताओं ने गांधीजी से मुलाकात की। *(Sardar Patel and other leaders met Gandhiji.)* Here, evidence provided to the MLN is:

   $NonIdentical$(सरदार पटेल, अन्य नेताओं)

2. **Consecutive actor mentions**: If any two actor mentions appear consecutively in a sentence, then these mentions are less likely to be coreferences of each other unless both are nominal mentions. E.g., उन्होंने किसानों की समस्याओं को समझा। *(He understood the difficulties of the farmers.)*

   $NonIdentical$(उन्होंने, किसानों)

   E.g., वे उनके अग्रज थे। *(They were his elder brothers.)*

   $NonIdentical$(वे, उनके)

3. **Noun modifiers**: If an actor mention is modifying (connected through a “nmod” dependency relation) another actor mention, then it is unlikely that these mentions are coreferences of each other. E.g., वे झवेरभाई पटेल की चौथी संतान थे। *(He was Jhaverbhai Patel’s fourth child.)*

   $NonIdentical$(झवेरभाई पटेल, चौथी संतान)

4. **Arguments of a single predicate**: If two actor mentions are arguments of a single verbal predicate (except copula or copula-like verbs), then it is unlikely that these mentions are coreferences of each other. In other words, such arguments represent two distinct *semantic roles* of a single verbal predicate. “Semantic Role Labelling” (SRL) is itself a difficult problem and there are no annotated SRL datasets for Hindi. Hence, to find arguments of verbs, we use dependency parsing as a surrogate for full-fledged SRL. An actor mention is said to be an argument of a verb if the verb is ancestor of the actor mention in dependency tree and there are no other actor mentions on the path to the verb. E.g., Figure 1 shows dependency tree for the sentence सरदार पटेल ने उन्हे कर न देने के लिये प्रेरित किया। *(Sardar Patel inspired them not to pay the taxes.)* Here, for the verbal predicate किया, सरदार पटेल (*agent*) and उन्हे (*patient*) are its arguments. For all pairs for such arguments of a single verb, we add a soft rule in the MLN indicating that the actor mentions in the pair do not refer to a single real life actor.

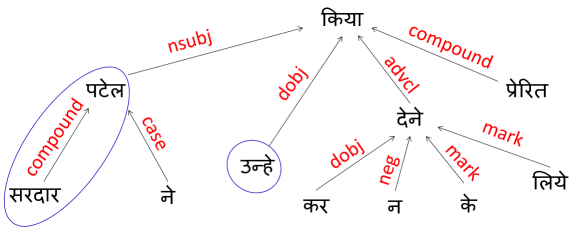

Figure 1: Dependency tree for the example sentence

$$10.0\ NonIdentical(\text{सरदार पटेल}, \text{उन्हे})$$

Antecedents: For each pronoun, we identify a certain "antecedent" nominal actor mention. The antecedent mention may precede the pronoun in the current sentence or be present in the previous sentence. To ensure that there is only one antecedent (y) for each pronoun (x), we add following rule with infinite weight:

$$IsAntecedent(x,y) \wedge (y \neq z) \Rightarrow \neg IsAntecedent(x,z).$$

Also, the pronoun (x) is likely to be coreferent of this antecedent (y). We incorporate this information with the following rule. The rule is a soft rule because it represents a weak assumption.

$$5.0\ IsAntecedent(x,y) \wedge PronounLike(x)$$
$$\wedge \neg PronounLike(y) \Rightarrow Coref(x,y)$$

Consider following text fragment from a narrative: किसान अंग्रेज सरकार से कर में छूट की मांग कर रहे थे। तब वे सरदार पटेल से मिले। *(The farmers were demanding a discount in the taxes from the British government. During that time they met Sardar Patel.)*

Here, for the pronoun वे *(they)*, we consider किसान *(farmers)* as its candidate antecedent which is its closest preceding and type-compatible nominal actor mention. Hence, for the given example, we add following evidence to the MLN:

$$IsAntecedent(\text{वे}, \text{किसान})$$

It is important to note here that even though अंग्रेज सरकार *(British government)* is the closest nominal actor mention for वे *(they)*, it gets skipped on grounds of type incompatibility (ORG vs PERSON). Also, this is an inter-sentence rule which enables us to establish inter-sentence coreference links.

In addition to pronouns, we identify antecedents for certain nominal actor mentions which are similar to "definite" mentions in English. These nominal actor mentions are generally preceded by a demonstrative pronoun. E.g., केवल तीन रियासतें छोड़कर उस लौह पुरुष ने सभी रियासतों को भारत में मिला दिया *(Leaving only three states, that Iron Man was able to merge all others into India.)* In this example sentence, लौह पुरुष *(Iron Man)* is a definite mention because it is preceded by a demonstrative pronoun उस *(that)*. Hence, we find other antecedent nominal mentions for लौह पुरुष, one of which is likely to be its coreferent.

3.3 Inference

First, an MLN is created for a given narrative using the above-mentioned rules and evidences. Then, we run marginal inference in this MLN for the query predicate $Coref(x,y)$. We select actor mention pairs (x,y) for which probability of $Coref(x,y)$ is above a certain threshold. These pairs represent coreference links. We further add additional links so as to ensure transitivity of coreferences and get final coreference clusters.

4 Experimentation Details

In this section, we explain the datasets, ground truth creation, the experimental setup, evaluation methodology and results with their analysis.

4.1 Datasets

We select four narratives from Hindi Wikipedia each corresponding to an important event or person in India's history and employ them as the datasets for our experiments. Table 3 describes basic statistics about the four datasets.

The raw Wikipedia text contained multiple issues which we corrected by performing some basic cleaning. The steps involved are as follows.

- Spelling correction: We observed multiple instances of incorrect spellings for words which we corrected manually by hand. For instance, the word नबाव *(nabaw)* was used a few times in place of intended word नवाब *(nawab)*.

- Spelling normalization: For difficult names like सिराजुदौला *(Sirajudaulah)* a single spelling was fixed and its multiple variations were normalized to the chosen canonical one. We also did this for names occurring with incorrect spellings. For example, we replaced the incorrect पी. वी. मेनन *(P. V. Menon)* with the correct utterance वी. पी. मेनन *(V. P. Menon)*

- Sentence splitting: The sentence end marker in Hindi is the purn viram sign । which is

Dataset	#Sentences	#Words	#Actor Mentions
sardar[3]	90	1459	305
plassey[4]	70	1214	243
shivaji[5]	70	1224	256
emergency[6]	56	1221	197

Table 3: Dataset details[7]

different from the English full stop (.) making it an unambiguous sentence end marker. Sentence splitting can thus be performed easily by splitting on “ । ” i.e. the purn viram followed by a space. However, at multiple places in the Wikipedia text sentences were either ending abruptly without an end marker or the next sentence started right behind the marker without any space. These cases were handled by splitting sentences manually.

- Wiki meta-data removal: The narratives were obtained directly from the Wikipedia articles available on the web and hence, Wikipedia meta-data such as reference numbers, bullets, inline links, etc. were present. These unwanted characters were also removed.

Apart from the four Hindi Wikipedia based datasets, we use another dataset IIITH Hindi Coreference (Dakwale, 2014; Mujadia, 2017) dataset. Unlike the earlier four datasets, coreference annotations were available for IIITH dataset. However, we had to manually revise the annotations due to the following reasons:

- The dataset contains annotations for non-actor mentions as well. E.g., फिल्म महोत्सव में प्रकाश झा की फिल्म अपहरण का भी प्रीमियर होना है। इस फिल्म में बिपाशा बसु ने भी बतौर अभिनेत्री काम किया है। *(In the film festival, Prakash Jha’s film Apaharan also has its premier. Bipasha Basu has acted in the movie as a lead actress.)* Hence, we discarded the non-actor mentions like फिल्म महोत्सव *(film festival)*, फिल्म *(film)* and अपहरण *(Apaharan)*.

- The dataset did not annotate entity types (PERSON, ORGANIZATION or LOCATION) for the mentions. Hence, we added actor types for all the actor mentions.

We manually revised annotations for 10 news articles out of the 275 news articles contained in the original dataset. These 10 new articles amounted to 156 sentences, 3412 words and 463 actor mentions.

4.2 Developing Ground Truth

An important part in the development of the ground truth is identification of actor mentions of the three types: named entities, generic NPs and pronouns for each dataset. The following tagging guidelines were set for guiding the actor mention identification.

- Tag all named entities occurring as separate noun phrases. For example, The phrase महात्मा गांधी *(Mahatma Gandhi)* needs to be tagged in महात्मा गांधी के विचारों से वे प्रेरित हुए । *(He was inspired by Mahatma Gandhi’s thoughts.)*

- Tag all named entities occurring as part of a noun phrase even if the whole noun phrase is not a PERSON, LOCATION or ORGANIZATION. For example, बारडोली *(Bardoli)* needs to be tagged in बारडोली सत्याग्रह में पटेल का महत्वपूर्ण योगदान रहा।*(Patel had a pivotal role in the Bardoli Satyagraha.)*

- Tag all generic NPs and pronouns which refer to PERSONs, LOCATIONs and ORGANIZATIONs.

- Tagging of certain generic NPs needs to be carried out depending on the context they occur in and the noun they modify. For example, in the sentence उप-प्रधानमंत्री सरदार पटेल ने भारत को जोडने का कठिन परिश्रम किया । *(Deputy Prime Minister Sardar Patel performed the difficult hardwork for uniting India.)*, the generic phrase उप-प्रधानमंत्री *(Deputy Prime Minister)* behaves as an adjectival modifier to सरदार पटेल *(Sardar Patel)* and hence, the whole phrase उप-प्रधानमंत्री सरदार पटेल *(Deputy Prime Minister Sardar Patel)* gets marked as a single mention. However in the sentence भारत के उप-प्रधानमंत्री सरदार पटेल ने रियासत विभाग का गठन कीया ।*(India’s Deputy Prime Minister, Sardar Patel constituted the States Ministry.)*, the generic phrase उप-प्रधानमंत्री would be tagged separately as it is

[3] https://hi.wikipedia.org/wiki/वल्लभ_भाई_पटेल

[4] https://hi.wikipedia.org/wiki/प्लासी_का_पहला_युद्ध

[5] https://hi.wikipedia.org/wiki/शिवाजी

[6] https://hi.wikipedia.org/wiki/आपातकाल_(भारत)

[7] The datasets and ground truth will be made available if the paper gets accepted.

being modified by the phrase भारत के *(India's)*. So, four segments should be marked from the sentence namely भारत *(India)*, उपप्रधानमंत्री *(Deputy Prime Minister)*, सरदार पटेल *(Sardar Patel)* and रियासत विभाग *(States Ministry)*.

Another important part of the ground truth is annotating a canonical mention for each actor mention to which it resolves to. Each canonical mention represents a cluster of actor mentions in which each mention is a coreference of other mentions. Also along with each actor mention, the type of the actor (PERSON, LOCATION or ORGANIZATION) is also specified by the annotator.

Four annotators were employed for creation of ground truth data. Each annotator tagged one dataset and then cross verified it with one other annotator. Tricky cases were discussed and resolved unanimously making sure the tagging guidelines were met and all annotators agree.

4.3 Experimental Setup

To tune the set of rules and their weights in the MLNs, we use the `sardar` dataset as a development dataset. The rest three datasets are only used for experimentation and reporting results. Also, as the main aim of the paper is to perform coreference resolution, we use the mentions identified in the text as a part of the ground truth as a starting point and run the MLN based algorithm to resolve these gold mentions.

To capture linguistic knowledge for predicates of the MLN we need the dependency parse of Hindi sentences. We used the Parsey Universal parser[8], available as part of the Google's Syntaxnet toolkit, which is a state-of-the-art open source dependency parser for 40 different languages.

For the implementation of the MLN, we use the open source MLN inference engine known as tuffy[9] (Niu et al., 2011). It supports marginal and MLE based inference. We use marginal inference in tuffy implemented through the MC-SAT algorithm with number of samples as 100. As this is an approximate inference, we run the inference for each narrative five times and report results averaged over the five runs.

Dataset	Setting	MUC	B^3	CEAFe	Avg
sardar	B1	74.63	59.33	50.57	61.51
	B2	70.55	69.67	63.57	67.93
	MLN	73.35	71.98	66.05	**70.46**
plassey	B1	72.3	53.03	47.93	57.75
	B2	62.36	61.39	62.74	62.16
	MLN	68.09	63.33	63.31	**64.91**
shivaji	B1	71.92	57.01	55.50	61.48
	B2	70.96	70.20	69.02	**70.06**
	MLN	71.07	70.00	65.88	68.98
emergency	B1	70.14	46.25	45.17	53.85
	B2	62.52	62.95	61.35	62.27
	MLN	62.93	63.62	62.83	**63.12**
IIIT-H	B1	67.53	53.51	42.84	54.62
	B2	59.24	50.42	38.81	49.49
	MLN	64.25	55.88	45.00	**55.04**

Table 4: F1 measures according to various metrics

4.4 Baselines and Evaluation

We developed following baseline approaches for Coreference Resolution of actor mentions:

1. **B1**: Here, a pair of actor mentions are said to be coreferences of each other if: i) they are lexically similar or ii) one of the mentions is pronoun and other is its type-compatible closest antecedent. Final coreference clusters are obtained by getting a transitive closure of such corefering pairs.

2. **B2**: This baseline uses the linguistic rules proposed by Patil et al. (2018). None of these rules capture any inter-sentence relation among actor mentions.

We evaluated the output of our MLN based system using three metrics widely used in the literature to report coreference resolution results. They are the MUC (Vilain et al., 1995), the B^3 (Bagga and Baldwin, 1998) and the CEAFe (Luo, 2005) metrics.

Table 4.4 reports the results obtained on the four datasets. It is important to note here that the approach proposed is unsupervised. Hence, the obtained accuracies are encouraging and entail further exploration.

4.5 Analysis

In Table 4, we report the comparative performance of our approach (MLN) with baselines. Our approach outperforms baseline B1 on all datasets and baseline B2 (Patil et al., 2018) on 4 out of 5 datasets. Even though baseline B2 is also based on MLN, our approach uses a richer set of rules such as:

[8] https://github.com/tensorflow/models/blob/master/research/syntaxnet/g3doc/universal.md

[9] http://i.stanford.edu/hazy/tuffy/

- Using the antecedent rule which enables us to link inter-sentential coreferences to a moderate extent. However, baseline B2 only uses intra-sentence rules.

- Finding antecedents not only for pronouns but also for definite mentions

- Using SRL-like predicate-arguments for identifying non-identical actor mentions

In general, English coreference resolution techniques exploit gender and number compatibility in addition to entity type compatibility. However, in Hindi, we observed that gender and number compatibility do not hold in large number of cases. As an example, the pronoun वे *(they / he)* may refer to the singular mention सरदार पटेल *(Sardar Patel)* or the plural mention किसानों *(farmers)* depending on the context. This is an example of a Hindi-specific phenomenon of using plurals to indicate respect (आदरार्थी बहुवचन). Similarly, the pronoun उन्हें *(him / her)* may refer to either the masculine mention संजय गांधी *(Sanjay Gandhi)* or the feminine mention इंदिरा गांधी *(Indira Gandhi)* depending on the context.

We also observed that considering only the nearest nominal actor mention as the antecedent, is not always correct. For example, in the sentence लालू प्रसाद यादव के साले सुभाष यादव ने उन पर गलत टिकट वितरण का आरोप लगाया *(Lalu Prasad Yadav's brother-in-law Subash Yadav accused him of incorrect candidacy allocation.)*, the pronoun उन *(him / her)* actually refers to लालू प्रसाद यादव *(Lalu Prasad Yadav)* and not सुभाष यादव *(Subash Yadav)* which is the closest type-compatible actor mention. This requires further exploration on using multiple preceding actor mentions as antecedents.

5 Conclusion and Future Work

In this paper, we focussed on resolving coreferences of actor mentions in Hindi narrative text. The proposed approach is linguistically grounded and uses Markov Logic Networks (MLN). The MLN framework proved to effective in representing various pieces of information characterizing linguistic knowledge relevant to coreference resolution. Unlike neural or other supervised approaches, our approach does not need a large amount of coreference annotated data. We also contributed four new datasets annotated with actor mentions and their coreferences. Our approach outperformed two baselines including a strong recent one developed for English narrative text.

In future, we plan to build an end-to-end system which first identifies actor mentions and then resolves coreferences among them. We also intend to analyze the robustness of our rules in this scenario of using predicted actor mentions.

References

S. Agarwal, M. Srivastava, P. Agarwal, and R. Sanyal. 2007. Anaphora resolution in hindi documents. In *2007 International Conference on Natural Language Processing and Knowledge Engineering*, pages 452–458.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653. Association for Computational Linguistics.

Praveen Dakwale. 2014. *Anaphora Resolution in Hindi*. Master's thesis, Language Technologies Research Center (LTRC), International Institute of Information Technology, Hyderabad - 500 032, INDIA.

Pedro M. Domingos and Daniel Lowd. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Kamlesh Dutta, Nupur Prakash, and Saroj Kaushik. 2008. Resolving Pronominal Anaphora in Hindi using Hobbs's algorithm. In *Web Journal of Formal Computation and Cognitive Linguistics*.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

J Hobbs. 1986. Readings in natural language processing. chapter Resolving Pronoun References, pages 339–352. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT/EMNLP 2005*, pages 25–32. Association for Computational Linguistics.

Vandan Mujadia. 2017. *Capturing and Resolving Entities and their Mentions in Discourse*. Master's thesis, Language Technologies Research Center (LTRC), International Institute of Information Technology, Hyderabad - 500 032, INDIA.

Vincent Ng. 2017. Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *AAAI, 2017*, pages 4877–4884.

Feng Niu, Christopher Ré, AnHai Doan, and Jude W. Shavlik. 2011. Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS. *PVLDB*, 4(6).

Sangameshwar Patil, Sachin Pawar, Swapnil Hingmire, Girish Palshikar, Vasudeva Varma, and Pushpak Bhattacharyya. 2018. Identification of alias links among participants in narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 63–68.

Rashmi Prasad and Michael Strube. 2000. Discourse salience and pronoun resolution in hindi. *Penn Working Papers in Linguistics*, 6(3):189–208.

Bhargav Uppalapu and Dipti Misra Sharma. 2009. Pronoun resolution for hindi. In *7th Discourse Anaphora and Anaphor Resolution Colloquium*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.