

End-to-End Relation Extraction using Markov Logic Networks

Sachin Pawar^{1,2}, Pushpak Bhattacharya^{2,3}, and Girish K. Palshikar¹
sachin7.p@tcs.com, pb@cse.iitb.ac.in, gk.palshikar@tcs.com

¹TCS Research, Tata Consultancy Services, Pune-411013, India

²Dept. of CSE, Indian Institute of Technology Bombay, Mumbai-400076, India

³Indian Institute of Technology Patna, Patna-801103, India

Abstract. The task of end-to-end relation extraction consists of two sub-tasks: i) identifying entity mentions along with their types and ii) recognizing semantic relations among the entity mention pairs. It has been shown that for better performance, it is necessary to address these two sub-tasks jointly [22, 13]. We propose an approach for simultaneous extraction of entity mentions and relations in a sentence, by using inference in Markov Logic Networks (MLN) [21]. We learn three different classifiers : i) local entity classifier, ii) local relation classifier and iii) “pipeline” relation classifier which uses predictions of the local entity classifier. Predictions of these classifiers may be inconsistent with each other. We represent these predictions along with some domain knowledge using weighted first-order logic rules in an MLN and perform joint inference over the MLN to obtain a global output with minimum inconsistencies. Experiments on the ACE (Automatic Content Extraction) 2004 dataset demonstrate that our approach of joint extraction using MLNs outperforms the baselines of individual classifiers. Our end-to-end relation extraction performance is better than 2 out of 3 previous results reported on the ACE 2004 dataset.

1 Introduction

Real world entities are referred in natural language sentences through *entity mentions* and these are often linked through meaningful *relations*. The task of end-to-end relation extraction consists of two sub-tasks: entity extraction and relation extraction. The sub-task of *entity extraction* deals with identifying entity mentions and determining their entity types. The other task of *relation extraction* deals with identifying whether any semantic relation exists between any two mentions in a sentence and also determining the relation type if it exists. In this paper, we refer to *entity extraction* and *relation extraction* tasks as defined by the Automatic Content Extraction (ACE) program [3] under the EDT (Entity Detection and Tracking) and RDC (Relation Detection and Characterization) tasks, respectively. ACE standard defined 7 entity types ¹: PER (person), ORG

¹ <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-edt-v4.2.6.pdf>

(organization), LOC (location), GPE (geo-political entity), FAC (facility), VEH (vehicle) and WEA (weapon). It also defined 7 coarse level relation types²: EMP-ORG (employment), PER-SOC (personal/social), PHYS (physical), GPE-AFF (GPE affiliation), OTHER-AFF (PER/ORG affiliation), ART (agent-artifact) and DISC (discourse).

Compared to the work (refer the surveys [18, 19]) in Named Entity Recognition (NER), there are relatively few attempts [4, 5, 13, 15] to address the more general entity extraction problem. NER extracts only named mentions (e.g. **John Smith**, **Walmart**) whereas entity extraction is expected to also identify common noun and pronoun mentions (e.g. **company**, **leader**, **it**, **they**) and their entity types. This task is more challenging than NER because entity type of mentions like **leader** or **they** may vary from sentence to sentence depending on which real life entity they are referring to in that sentence. For example, entity type of **leader** would be PER in the sentence **John Smith was elected as the leader of the Socialist Party** whereas its entity type would be ORG in the sentence **Pepsi is a market leader in its segment**.

There has been a lot of work for relation extraction like Zhou et al. [6], Jiang and Zhai [10], Bunescu and Mooney [1] and Qian et al. [20]. All of these approaches assume that the boundaries and the types of entity mentions are already known. Several features based on this information are used for relation prediction. In order to use such relation extraction systems, there should be separate entity extraction system whose output acts as an input for relation extraction. In such a “pipeline” method, the errors are propagated from first phase (entity extraction) to second phase (relation extraction) affecting the overall relation extraction performance. Another major disadvantage of the “pipeline” method is that it facilitates only one-way *information flow*, i.e. the knowledge about entities is used for relation extraction but not vice versa. However, the knowledge about relations can help in correcting some entity extraction errors.

In order to overcome these problems, we propose an approach which uses inference in Markov Logic Networks (MLN) for simultaneous extraction of entities and relations in a sentence. This approach facilitates two-way *information flow*. MLNs combine first-order logic and probabilistic graphical models in a single representation. An MLN contains a set of first-order logic rules, and each rule is associated with a weight. The fewer rules a world violates, the more probable it is. Also, higher the weight of a rule, greater is the probability of a world that satisfies the rule compared to the one that does not. In our approach, three separate classifiers are learned: a local entity classifier, a local relation classifier and a “pipeline” relation classifier which uses predictions of the local entity classifier. Predictions of these classifiers along with other domain knowledge are represented using weighted first-order logic rules in an MLN. Joint inference over this MLN is then performed to get a final output with least possible contradictions or inconsistencies among the individual classifiers.

The specific contributions of this work are : i) a novel approach for joint extraction of entity mentions and relations using inference in MLNs and ii) easy

² <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-rdc-v4.3.2.PDF>

and compact representation of the domain knowledge using first-order logic rules in MLNs. The rest of the paper is organized as follows. Section 2 describes some background and necessary building blocks for our approach. Section 3 describes our approach in detail and Section 4 describes the working of our approach through an example. Experimental results are presented in Section 5. Related work is then described briefly in Section 6. Finally we conclude in Section 7 with brief discussion about the future work.

2 Building Blocks for Our Approach

2.1 Markov Logic Networks

Markov Logic Networks (MLN) which were proposed by Richardson and Domingos [21], combine first-order logic and probabilistic graphical models in a single representation. Formally, a Markov Logic Network L is defined as a set of pairs (F_i, w_i) , where each F_i is a formula in first-order logic with a real weight w_i . Along with a finite set of constants $C = \{C_1, C_2, \dots, C_{|C|}\}$, it defines a Markov Network $M_{L,C}$ as follows:

1. $M_{L,C}$ contains one binary node for each possible grounding of each predicate appearing in L . The value of the node is 1 if the ground atom is true, and 0 otherwise.
2. $M_{L,C}$ contains one feature for each possible grounding of each formula F_i in L . The value of this feature is 1 if the ground formula is true, and 0 otherwise. The weight of the feature is the w_i associated with F_i in L .

The probability distribution of random variable X over possible worlds x specified by Markov Network $M_{L,C}$ is given by,

$$P(X = x) = \frac{1}{Z} \exp \left(\sum_i w_i n_i(x) \right) \quad (1)$$

where $n_i(x)$ is the number of true groundings of F_i in x and Z is the partition function. MLN can be used to find probability of a formula (say F_1) being true, given some other formula (say F_2) is true.

$$P(F_1|F_2, M_{L,C}) = \frac{P(F_1 \wedge F_2|M_{L,C})}{P(F_2|M_{L,C})} = \frac{\sum_{x \in X_{F_1} \cap X_{F_2}} P(X = x|M_{L,C})}{\sum_{x \in X_{F_1}} P(X = x|M_{L,C})}$$

where X_{F_i} represents the set of worlds where F_i holds and $P(X = x|M_{L,C})$ is computed using the equation 1.

2.2 Identifying Entity Mention Candidates

It is necessary to first identify the span (or boundaries) of each entity mention ³ in a given sentence. We model this as a sequence labelling problem. A sentence

³ We consider the ‘‘head’’ extent of a mention defined by ACE standard as the entity mention so that all the valid entity mentions are always non-overlapping.

is a sequence of words and each word in a sentence is assigned a label indicating whether that word belongs to any entity mention or not. We use BIO encoding for this purpose.

- **O**: Label for the words which are not part of any entity mention
- **B**: Label for the first word of entity mentions
- **I**: Label for the subsequent words (except the first word) of entity mentions

We employ the Conditional Random Field (CRF) model [12], which is trained in a supervised manner. Given any new sentence, we use the trained CRF model to predict the 2 most probable label sequences as follows:

S_1 :A/O Palestinian/B Council/B member/B says/O anger/O is/O growing/O ./O
 S_2 :A/O Palestinian/B Council/I member/B says/O anger/O is/O growing/O ./O

In this sentence, entity mention candidates from the topmost sequence are **Palestinian**, **Council** and **member**. Entity mention candidate **Palestinian Council** is generated from the second sequence. Generally, the candidates generated from the most probable sequence are more likely to be valid entity mentions. The candidates generated from the second most probable sequence are considered valid entity mentions only if they satisfy certain constraints. These constraints are applied in the form of first-order logic rules in MLNs and will be explained later. A special entity type *NONE* is assigned to a candidate entity mention if it is an invalid entity mention.

2.3 Local Entity Classifier

The local entity classifier is used to predict the most probable entity type for each candidate entity mention in a given sentence. This classifier is referred to as “local” as it takes an independent decision for each entity mention irrespective of its relation with other mentions. A Maximum Entropy Classifier is trained in a supervised manner which captures the characteristics of each entity mention E using following features:

1. **Lexical Features**: Head word and other words in E , words preceding and succeeding E in the sentence
2. **Syntactic Features**: POS tags of the head word and other words in E , POS tags of the words preceding and succeeding E , parent of head word of E in the dependency tree and also the dependency relation with the parent
3. **Semantic Features**: WordNet category (if any) of the head word of E . Some specific synsets in the WordNet (e.g. **person**, **location**, **vehicle**) are marked as possible “categories” and if any word is direct or indirect hypernym of such synsets, it is said to be falling in the corresponding “category”.

As this classifier is trained using only the valid entity mentions in the training data, it always predicts one of the 7 ACE entity types and never predicts the *NONE* type.

2.4 Local Relation Classifier

The local relation classifier is used to predict the most probable relation type for each pair of candidate entity mentions in a given sentence. This classifier is referred to as “local” as it takes an independent decision for each pair of entity mentions irrespective of their entity types.

In addition to ACE 2004 relation types, it considers two special relation types “NULL” (indicating that no semantic relation holds) and “IDN” (representing intra-sentence co-references). In the sentence *Pepsi is a market leader*, the entity mentions *leader* and *Pepsi* are co-references and hence we add the IDN relation between these mentions. With the help of IDN (identity) relation type, information about intra-sentence co-references can be incorporated in a principled way without using an external co-reference resolution system. Also, more number of entity mentions get involved in at least one relation, resulting in better entity extraction performance. For example, in the ACE 2004 dataset, there are 22718 entity mentions and 4328 relation instances resulting in only 7604 entity mentions involved in at least one relation. Considering the IDN relation, number of relation instances increases to 12060 covering 14930 entity mentions.

A Maximum Entropy Classifier is used which captures the characteristics of each entity mention pair (E_1, E_2) with the help of following features:

1. **Lexical Features:** Head words and other words of E_1 & E_2 , words preceding and succeeding E_1 & E_2 in the sentence
2. **Syntactic Features:** POS tags of the head word and other words in E_1 & E_2 , POS tags of the words preceding and succeeding E_1 & E_2 , parents of head words of E_1 & E_2 in the dependency tree and also the dependency relations with the parents, path connecting E_1 & E_2 in the dependency tree, their common ancestor in the dependency tree
3. **Semantic Features:** WordNet categories (if any) of the head words of E_1 & E_2 , the common ancestor and other words on the path connecting E_1 & E_2 in the dependency tree of the sentence, syntactico-semantic structures identified in Chan and Roth [2].

2.5 Pipeline Relation Classifier

Unlike the local relation classifier, the “pipeline” relation classifier is dependent on the output of the local entity classifier. It uses following features in addition to the features used by the local relation classifier.

1. Entity types of E_1 and E_2 as predicted by the local entity classifier
2. Concatenation of entity types of E_1 and E_2
3. A binary feature indicating whether the types of E_1 and E_2 are same or not.

This classifier is referred to as a “pipeline” classifier because of unidirectional *information flow*. In other words, the knowledge about types of entity mentions is used by the relation classifier but not vice versa.

3 Joint Extraction using Inference in MLNs

3.1 Motivation

As described in the previous section, we have 3 classifiers producing various predictions about entity types and relation types. These decisions may be inconsistent, i.e. relation type predicted by the local relation classifier may not be compatible with the entity types predicted by the local entity classifiers. Also, there may be contradiction in predictions of local relation classifier and “pipeline” relation classifier. Our aim is to take predictions of these classifiers as input and make a global prediction which minimizes such inconsistencies. MLN provides a perfect framework for this, where we can represent predictions of individual classifiers as first-order logic rules where weights of these rules are proportional to the prediction probabilities (soft constraints). Also, the consistency constraints among the relation types and entity types can be represented in the form of first-order logic rules with infinite weights (hard constraints). Now, the inference in such an MLN will provide a globally consistent output with maximum weighted satisfiability of the rules. The detailed explanation is provided in subsequent sections about how the first-order logic rules are created and how the corresponding weights are set.

3.2 Domains and Predicates

We specify one MLN for a sentence, i.e. for all candidate entity mentions and possible relation instances in a sentence. The software package used for inference in MLN is Alchemy⁴. We define 3 domains : *entity*, *etype* and *rtype*. The *entity* domain represents entity mentions where a unique ID is assigned to each entity mention. It is specified as follows in Alchemy for a sentence with n entity mentions having IDs from 1 to n :

$$entity = \{1, 2, \dots, n\}$$

The next domain *etype* represents the set of all possible entity types and another domain *rtype* represents the set of all possible relation types. These domains are specified in Alchemy as follows:

$$etype = \{PER, ORG, LOC, GPE, WEA, FAC, VEH, NONE\}$$

$$rtype = \{EMPORG, GPEAFF, OTHERAFF, PERSOC, PHYS, ART, NULL, IDN\}$$

We define following predicates which are used for writing various first-order logic rules. The arguments for these predicates come from the above domains.

1. $ET(entity, etype)$: $ET(i, E)$ is true only when entity type of the entity mention i is equal to E . It is true for one and only one entity type. It represents the entity type prediction of the local entity classifier and used as an *evidence* during inference.

⁴ <http://alchemy.cs.washington.edu/>

2. $RTP(entity, entity, rtype)$: $RTP(i, j, R)$ is true only when type of relation between entity mentions i and j is equal to R . It is true for one and only one relation type. It represents relation type prediction of “pipeline” relation classifier. It is also used as an *evidence*.
3. $RTL(entity, entity, rtype)$: Similar to RTP but represents relation type prediction of local relation classifier.
4. $ETFinal(entity, etype)$: Similar to ET but represents global entity type prediction and is used as a *query* predicate during inference.
5. $RTFinal(entity, entity, rtype)$: Similar to RTP but represents global relation type prediction and is used as a *query* predicate.

During the inference in MLN, the probabilities of all possible groundings of *query* predicates are computed, conditioned on the specific groundings of the *evidence* predicates.

3.3 Generic Rules

Although one MLN is created for each sentence, some first-order rules are common and they are added to MLNs of all the sentences. We refer to these rules as *Generic Rules*. These rules represent some universal truths about the domain and hence the weight associated with each of these rules is set to *infinity*. In other words, any world that violates any of these rules, is practically impossible. These rules provide an easy and effective way of incorporating the domain knowledge about entity types and relation types. For each valid combination of relation type and entity type of one of its argument, we write rules to constrain the possible entity types for the other argument. Such rules can be easily devised by going through the ACE 2004 labelling guidelines. Following are some representative examples⁵. Note that the variables x, y are universally quantified at the outermost level.

1. If there is an *EMPORG* relation between two entity mentions and entity type of any mention is *PER*, then entity type of other mention can only be one of : *ORG* or *GPE*.

$$RTFinal(x, y, EMPORG) \wedge ETFinal(x, PER) \Rightarrow (ETFinal(y, ORG) \vee ETFinal(y, GPE)).$$

$$RTFinal(x, y, EMPORG) \wedge ETFinal(y, PER) \Rightarrow (ETFinal(x, ORG) \vee ETFinal(x, GPE)).$$

2. For the “identity” relation type *IDN*, the constraint is that the entity types of both the mentions should be same.

$$RTFinal(x, y, IDN) \wedge ETFinal(x, z) \Rightarrow ETFinal(y, z).$$

$$RTFinal(x, y, IDN) \wedge ETFinal(y, z) \Rightarrow ETFinal(x, z).$$

⁵ All the rules can’t be listed because of the space constraints.

3.4 Sentence-specific Rules

These rules are specific to each sentence and represent the predictions by the individual baseline classifiers. Unlike the *Generic Rules*, these rules are added with finite weights.

Weight Assignment Strategies: In order to learn the weights of various first-order logic rules, historical examples of predictions of 3 base classifiers along with gold-standard predictions would be required. Instead we chose to compute these weights by using some functions of the corresponding prediction probabilities. The work by Jain [9] discussed various ways of weight assignments to represent knowledge in MLNs. In another work, Heckmann et al. [8] adjusted the rule weights experimentally for citation segmentation using MLNs. On the similar lines, following two strategies are adopted for weight assignments.

1. **Log of Odds Ratio (LOR):** Richardson and Domingos [21] states that the weight of a formula F is log odds between a world where F is true and a world where F is false. For a prediction with probability p , we set the weight of corresponding formula as $\log\left(\frac{p}{1-p}\right)$. Here, the penalty for violating any formula will increase logarithmically with its probability.
2. **Constant Multiplier (CM):** As per this strategy, for a prediction with probability p , we set the weight of corresponding formula as $K \cdot p$. Here, the penalty for violating any formula will increase linearly with its probability. We have used $K = 10$ in all our experiments.

Rules induced by the Local Entity Classifier: For each candidate entity mention, the entity type predicted by the local entity classifier acts as an *evidence* for the MLN inference. The classifier also assigns some probability to each possible entity type. For each entity mention id i , for each possible entity type E , following rule is added with the weight proportional to the probability of prediction.

$$ET(i, E_{max}) \Leftrightarrow ETFinal(i, E)$$

Here, E_{max} is the entity type predicted by the local entity classifier. The weights assigned to this rule as per above strategies would be $\log\left(\frac{P_e(E|i)}{1-P_e(E|i)}\right)$ and $K \cdot P_e(E|i)$, where $P_e(E|i)$ is the probability assigned to entity type E for the entity mention id i by the local entity classifier.

Rules induced by the Pipeline Relation Classifier: For each pair of entity mentions, the relation type predicted by the “pipeline” classifier acts as an *evidence* for the MLN inference. For each pair of candidate entity mentions (i, j) , for each possible relation type R , following rule is added with the weight proportional to the probability of prediction.

$$RTP(i, j, R_{max}) \Leftrightarrow RTFinal(i, j, R)$$

Here, R_{max} is the relation type predicted by the “pipeline” relation classifier. The weights assigned to this rule would be $\log\left(\frac{wt_p \cdot P_r^P(R|i, j)}{1 - P_r^P(R|i, j)}\right)$ and $K \cdot P_r^P(R|i, j) \cdot wt_p$, where $P_r^P(R|i, j)$ is the probability assigned to the relation type R for

the pair (i, j) by the “pipeline” relation classifier. And wt_p is the reliability of prediction of “pipeline” classifier, which indicates how confident the local entity classifier was in predicting entity types for entity mentions i and j . We set $wt_p = P_e(E_{max}^i|i) \cdot P_e(E_{max}^j|j)$.

Rules induced by the Local Relation Classifier: For each pair of candidate entity mentions, the relation type predicted by the local classifier acts as an *evidence* for the MLN inference. For each pair entity mentions (i, j) , for each possible relation type R , following rule is added with the weight proportional to the probability of prediction.

$$RTL(i, j, R_{max}^L) \Leftrightarrow RTFinal(i, j, R)$$

Here, R_{max}^L is the relation type predicted by the local relation classifier. The weights assigned to this rule would be $\log\left(\frac{P_r^L(R|i, j)}{1 - P_r^L(R|i, j)}\right)$ and $K \cdot P_r^L(R|i, j)$, where $P_r^L(R|i, j)$ is the probability assigned to the relation type R for the pair (i, j) by the local relation classifier.

Rules for identifying valid/invalid entity mentions: We generate candidate entity mentions using top 2 most probable BIO sequences. In general, we have a high confidence that candidate mentions from the topmost sequence are valid and have a lower confidence for candidates from the second sequence. This intuition is captured by addition of following rules. For each candidate i from the topmost sequence, we add $!ETFinal(i, NONE)$ with the weight $\log\left(\frac{p}{1-p}\right)$ or $K \cdot p$, based on the weighing strategy employed. Also for each candidate i from the second sequence, we add $ETFinal(i, NONE)$ with the weight $\log\left(\frac{1-p}{p}\right)$ or $K \cdot (1 - p)$. In both the cases, p is the highest probability for any entity type predicted for that mention by the local entity classifier. As we are generating candidate entity mentions by using top 2 most probable BIO sequences, there may be some overlapping entity mentions. For each pair of such overlapping candidate entity mentions (say i and j), following rules are added so that at most one of them is a valid entity mention.

$$!ETFinal(i, NONE) \Rightarrow ETFinal(j, NONE).$$

$$!ETFinal(j, NONE) \Rightarrow ETFinal(i, NONE).$$

We assume candidate mentions generated from the second BIO sequence to be valid, only if they are involved in some valid relation other than *NULL*. Also, an invalid entity mention should not be involved in any non *NULL* relation with any other mention. To ensure this desired consistency, following rules are added for each pair of candidate mentions (i, j) where one of them (say i) is generated from second BIO sequence.

$$!RTFinal(i, j, NULL) \Rightarrow !ETFinal(i, NONE).$$

$$ETFinal(i, NONE) \Rightarrow RTFinal(i, j, NULL).$$

After the inference, if the probability of $ETFinal(i, NONE)$ is the highest for any candidate mention i , then it is identified as an invalid mention. And because

of above rules ensuring consistency, such mentions are never involved in any non *NULL* relation.

3.5 Additional Semantic Rules

We explored the possibility of incorporating some domain knowledge by exploiting the easy and effective representability of the first-order logic. In order to incorporate the additional rules, we define following new predicates:

1. $CONS(entity, entity) : CONS(i, j)$ is true only when there is no other entity mention occurring in between the mentions i and j in a sentence.
2. $CONJ(entity, entity) : CONJ(i, j)$ is true only when there is a conjunction (i.e. connected through the dependency relations “conj:and” or “conj:or” in the dependency tree) between the two mentions i and j .

Using the knowledge of conjunctions: When two entity mentions are connected through a conjunction (like **and**, **or**) and one of them is connected to a third entity mention with *PHYS* (i.e. located at) relation, then the other entity mention is also very likely to be connected to the third mention with *PHYS* relation. E.g. in the sentence fragment **troops in Israel and Syria**, a *PHYS* relation between **troops** and **Israel** implies another *PHYS* relation between **troops** and **Syria**. To incorporate this knowledge, following generic rules are added in MLNs of all sentences.

$$RTFinal(x, y, PHYS) \wedge ((CONJ(y, z) \wedge CONS(y, z)) \vee (CONJ(z, y) \wedge CONS(z, y))) \\ \wedge ET(y, t) \wedge ET(z, t) \Rightarrow RTFinal(x, z, PHYS).$$

$$RTFinal(x, y, PHYS) \wedge ((CONJ(w, x) \wedge CONS(w, x)) \vee (CONJ(x, w) \wedge CONS(x, w))) \\ \wedge ET(w, t) \wedge ET(x, t) \Rightarrow RTFinal(w, y, PHYS).$$

Linking entity mentions with same types: The entity mentions linked through certain dependency relations tend to share the same entity type. E.g. in the sentence fragment **companies such as Nielsen**, the mentions **companies** and **Nielsen** are very likely to have the same entity type. This is one of the Hearst patterns [7] to automatically identify hyponyms from text. If entity mentions i and j follow such a pattern, we add following rule to their sentence’s MLN : $ETFinal(i, x) \Leftrightarrow ETFinal(j, x)$.

Using knowledge about relation types: If an entity mention of type *PER* is involved in a *EMPORG* relation, then it is highly unlikely that the same person will be connected to any other mention with the *EMPORG* relation. This is because any person can have at most one employer mentioned in a single sentence. To impose this constraint, we add following rule.

$$RTFinal(x, y, EMPORG) \wedge (y \neq z) \wedge \neg RTFinal(y, z, IDN) \wedge \neg RTFinal(z, y, IDN) \\ \wedge ETFinal(x, PER) \Rightarrow \neg RTFinal(x, z, EMPORG) \wedge \neg RTFinal(z, x, EMPORG).$$

3.6 Joint Inference

As described above, an MLN is created for a sentence using some *Generic Rules* with infinite weights and some sentence-specific rules. Given such an MLN, we are interested to know the most probable groundings of the *query* predicates given some specific groundings of *evidence* predicates. In our case, *ETF_{Final}* and *RTF_{Final}* are the *query* predicates and *ET*, *RTP*, *RTL*, *CONS* and *CONJ* are the *evidence* predicates. Inference over this MLN gives the probability of each possible grounding of the *query* predicates, conditioned on the given values of the *evidence* predicates. We used the default inference algorithm in Alchemy named “Lifted Belief Propagation” [26]. For each candidate entity mention *i*, grounding of the predicate *ETF_{Final}(i, E)* with the highest probability is chosen and corresponding value of *E* is its final entity type except the case when *E* = *NONE*. In that case, we do not identify the corresponding candidate mentions as a valid entity mention. Similarly, for each entity mention pair (*i, j*), grounding of the predicate *RTF_{Final}(i, j, R)* with the highest probability is chosen and corresponding *R* value is its final relation type.

4 Example

In this section, we describe an example sentence where the joint inference helps in correcting the prediction errors by the individual classifiers. Consider the sentence from the ACE 2004 dataset: **she is the new chair of the black caucus**. In order to identify the candidate entity mentions, top 2 label sequences predicted by the CRF model are considered.

1. she/B is/O the/O new/O chair/O of/O the/O black/O caucus/B ./O
2. she/B is/O the/O new/O chair/B of/O the/O black/O caucus/B ./O

Table 1 shows all the candidate entity mentions identified along with their IDs and predictions of the local entity classifier. It can be observed that mention ID

Table 1. Candidate entity mentions identified in the example sentence

ID	Entity Mention	From First BIO Sequence?	Predicted Type	Actual Type
1	she	Yes	PER	PER
2	chair	No	PER	PER
3	caucus	Yes	PER	ORG

2 is generated from the second best BIO sequence and hence will be considered a valid mention only if it is involved in a relation with some other mention. Moreover, the entity type predicted for the mention ID 3 (**caucus**) is incorrect. This error propagates to the relation classification with “pipeline” classifier predicting relation between **chair** and **caucus** to be IDN instead of EMP-ORG. But the local classifier predicts the correct relation type EMP-ORG for this pair as it is

not using the entity type features. The first-order logic rules for this sentence’s MLN are shown in the Table 2. The LOR (log of odds ratio) weights assignment strategy is used. In case of soft constraints, the number preceding each rule indicates its weight. No weight is explicitly specified for the hard constraints and they always end with a period.

Table 2. First-order logic rules for the MLN of example sentence

Rules induced by the local entity classifier	Rules for identifying valid/invalid entity mentions
6.13 $ET(1,PER) \Leftrightarrow ETFinal(1,PER)$	6.13 $\neg ETFinal(1,NONE)$
-0.93 $ET(2,PER) \Leftrightarrow ETFinal(2,LOC)$	0.71 $ETFinal(2,NONE)$
-0.89 $ET(2,PER) \Leftrightarrow ETFinal(2,ORG)$	0.15 $\neg ETFinal(3,NONE)$
-0.71 $ET(2,PER) \Leftrightarrow ETFinal(2,PER)$	$ETFinal(2,NONE) \Rightarrow RTFfinal(1,2,NULL)$.
-0.53 $ET(3,PER) \Leftrightarrow ETFinal(3,ORG)$	$\neg RTFfinal(1,2,NULL) \Rightarrow \neg ETFinal(2,NONE)$.
0.15 $ET(3,PER) \Leftrightarrow ETFinal(3,PER)$	$ETFinal(2,NONE) \Rightarrow RTFfinal(2,3,NULL)$.
	$\neg RTFfinal(2,3,NULL) \Rightarrow \neg ETFinal(2,NONE)$.
Rules induced by the local and pipeline relation classifiers	
3.37 $RTL(1,2,IDN) \Leftrightarrow RTFfinal(1,2,IDN)$	
2.99 $RTP(1,2,IDN) \Leftrightarrow RTFfinal(1,2,IDN)$	
1.52 $RTL(1,3,NULL) \Leftrightarrow RTFfinal(1,3,NULL)$	
-1.66 $RTL(1,3,NULL) \Leftrightarrow RTFfinal(1,3,IDN)$	
0.35 $RTP(1,3,NULL) \Leftrightarrow RTFfinal(1,3,NULL)$	
-1.63 $RTP(1,3,NULL) \Leftrightarrow RTFfinal(1,3,IDN)$	
-1.80 $RTL(2,3,EMPORG) \Leftrightarrow RTFfinal(2,3,PHYS)$	
-1.09 $RTL(2,3,EMPORG) \Leftrightarrow RTFfinal(2,3,IDN)$	
0.24 $RTL(2,3,EMPORG) \Leftrightarrow RTFfinal(2,3,EMPORG)$	
-0.46 $RTP(2,3,IDN) \Leftrightarrow RTFfinal(2,3,IDN)$	

Table 3. MLN inference output for entity types

she (ID 1)	chair (ID 2)	caucus (ID 3)
$ETFinal(1,PER) = \mathbf{0.99}$	$ETFinal(2,PER) = \mathbf{0.92}$	$ETFinal(3,PER) = 0.35$
$ETFinal(1,GPE) = 0.01$	$ETFinal(3,GPE) = 0.03$	$ETFinal(3,ORG) = \mathbf{0.39}$
	$ETFinal(2,GPE) = 0.02$	$ETFinal(3,NONE) = 0.14$
	$ETFinal(2,NONE) = 0.01$	$ETFinal(3,FAC) = 0.03$

The joint inference combines the evidence from the above three classifiers and generates a globally consistent output. The outputs for the query predicates $ETFinal$ and $RTFfinal$ are shown in the Tables 3 and 4, respectively. The predicate groundings which have negligible probability are not shown. Here, it can be observed that the entity mention **chair** (ID 2) has been correctly identified as a valid mention and the type of entity mention **caucus** has been correctly predicted as **ORG**. Also the correct relation type of **EMP-ORG** between **chair** and **caucus** has been chosen as the global prediction.

Table 4. MLN inference output for relation types

(she, chair)	(she, caucus)
$RTFinal(1, 2, IDN) = \mathbf{0.92}$	$RTFinal(1, 3, EMPORG) = 0.02$
$RTFinal(1, 2, PHYS) = 0.01$	$RTFinal(1, 3, PERSOC) = 0.02$
$RTFinal(1, 2, ART) = 0.01$	$RTFinal(1, 3, OTHERAFF) = 0.02$
$RTFinal(1, 2, OTHERAFF) = 0.02$	$RTFinal(1, 3, NULL) = \mathbf{0.90}$
$RTFinal(1, 2, NULL) = 0.01$	$RTFinal(1, 3, IDN) = 0.02$
(chair, caucus)	
$RTFinal(2, 3, EMPORG) = \mathbf{0.33}$	
$RTFinal(2, 3, PHYS) = 0.10$	
$RTFinal(2, 3, GPEAFF) = 0.10$	
$RTFinal(2, 3, OTHERAFF) = 0.09$	
$RTFinal(2, 3, IDN) = 0.20$	

5 Experimental Analysis

In order to demonstrate the effectiveness of our approach, we compare its performance with other approaches which have reported their results for end-to-end relation extraction on ACE 2004 dataset ⁶. For fair comparison, we follow the same assumptions made by Chan and Roth [2] and Li and Ji [13], i.e. ignoring the DISC relation, not treating implicit relations as false positives and using coarse entity and relation types. All the results are obtained by 5-fold cross-validation on ACE-2004 data. Note that the actual folds used by each algorithm may differ.

Table 5. Results on the ACE 2004 dataset (Micro-averaged, 5-fold cross-validation)

Approach	Entity Extraction			Relation Extraction			Entity+Relation		
	P	R	F	P	R	F	P	R	F
Local Classifiers	80.9	77.6	79.2	53.2	43.9	48.1	46.2	38.1	41.8
Pipeline Classifier				53.3	46.4	49.6	48.7	42.5	45.4
Chan and Roth [2]				42.9	38.9	40.8			
Li and Ji [13]	83.5	76.2	79.7	64.7	38.5	48.3	60.8	36.1	45.3
Miwa and Bansal [16]	83.3	79.2	81.2				56.1	40.8	47.2
MLN (LOR)	79.3	79.9	79.6	56.2	45.2	50.1	50.6	40.8	45.2
MLN (LOR)+Rules	79.3	80.0	79.6	56.6	45.1	50.2	51.0	40.6	45.2
MLN (CM)	78.9	80.1	79.5	57.2	45.2	50.5	51.6	40.8	45.6
MLN (CM)+Rules	79.0	80.1	79.5	57.9	45.6	51.0	52.4	41.3	46.2

Comparative performances of all the approaches are shown in the table 5. A true positive for the task of entity extraction means that an entity mention has been correctly identified as the valid mention and also its type has been identified correctly. A true positive for the task of relation extraction means that for a pair of valid entity mentions, its relation type (except for special relation types

⁶ We have not yet acquired a more recent ACE 2005 dataset

NULL and *IDN*) has been identified correctly. For entity+relation extraction, a stricter criteria is used where a true positive means that for a pair of valid entity mentions, not only its relation type is identified correctly but types of both the mentions are also identified correctly. Even if any one of these predictions is incorrect, we consider it as a false positive for the predicted combination of entity types and relation type and also as a false negative for the true combination of entity types and relation type.

It can be observed that MLN inference with CM (Constant Multiplier) weights assignment strategy performs better than the LOR (Log of Odds Ratio) in case of relation extraction whereas for entity extraction LOR strategy is better. Addition of semantic rules (discussed in the Section 3.5) results in better performance for both the strategies. Also, we can observe that MLN (CM) with semantic rules comfortably outperforms the individual classifiers: local entity classifier, local relation classifier and “pipeline” relation classifier. In case of end-to-end relation extraction, our approach outperforms the approaches of Chan and Roth [2] and Li and Ji [13] on the ACE 2004 dataset and also achieves a comparable performance as compared to Miwa and Bansal [16]. We also achieve comparable performance in case of entity extraction as compared to Li and Ji [13] but underperform in comparison with Miwa and Bansal [16].

6 Related Work

Previous work on joint extraction of entities and relations can be broadly classified into 5 categories : i) Integer Linear Programming (ILP) based approaches [22, 24], ii) Probabilistic Graphical Models [23, 25], iii) Card-pyramid parsing [11], iv) Structured Prediction [13, 14, 17] and v) Recurrent Neural Network (RNN) based model [16]. Our approach is similar to ILP based approaches, but we use MLNs for joint inference which provide much better representation to incorporate complex domain knowledge as compared to ILP. For example, the rules defined in the section 3.5 are quite easy to incorporate using first-order logic but the same would be cumbersome in ILP. The approaches by Singh et al. [25] and Li and Ji [13] not only carry out joint “inference” but also create a joint “model” where the parameters for both the tasks are learned jointly.

Zhang et al. [27] used Markov Logic rules to perform *Ontological Smoothing*. The concept of *Ontological Smoothing* is to find a mapping from a user-specified target relation to a background knowledge base. This mapping is then used to generate extra training data for distant supervision. Similar to our approach, they also use Markov logic rules to ensure consistency between relation types and entity types. One major difference is that the relation types used by them were quite specific and not as general as ACE 2004 relation types. Zhu et al. [28] also used MLNs but they addressed a relation extraction problem which is bit different from the ACE 2004 RDC task. It requires the explicit mention of relation in the form of words other than the words inside entity mentions. This is not always true for ACE 2004 relations. For example, EMP-ORG relation holds be-

tween Indian and soldiers in the sentence Indian soldiers attacked the terrorists.

7 Conclusion and Future Work

We described the problem of end-to-end relation extraction and the need to jointly address its sub-tasks of entity and relation extraction. We proposed a new approach for joint extraction of entity mentions and relations at the sentence level, which uses joint inference in Markov Logic Networks (MLN). We described in detail about the domains, predicates and first-order logic rules used to create an MLN for a sentence. We also explored how the effective representability of first-order logic can be used to incorporate various semantic rules and domain knowledge. Finally, we demonstrated better than the state-of-the-art end-to-end relation extraction performance on the standard dataset of ACE 2004.

In future, we plan to analyze the two weights assignment strategies (CM and LOR) in detail and develop deeper understanding of pros and cons of each one. Also, we have tried only a small number of additional semantic rules. In future, we wish to take advantage of the first-order logic framework to incorporate deeper semantic knowledge. Another important direction to explore is about learning the weights of first-order logic rules automatically.

References

1. Bunescu, R.C., Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 724–731. ACL (2005)
2. Chan, Y.S., Roth, D.: Exploiting syntactico-semantic structures for relation extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 551–560. ACL (2011)
3. Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S., Weischedel, R.M.: The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In: LREC. vol. 2, p. 1 (2004)
4. Florian, R., Jing, H., Kambhatla, N., Zitouni, I.: Factorizing complex models: A case study in mention detection. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. pp. 473–480. ACL (2006)
5. Florian, R., Pitrelli, J.F., Roukos, S., Zitouni, I.: Improving mention detection robustness to noisy input. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 335–345. ACL (2010)
6. GuoDong, Z., Jian, S., Jie, Z., Min, Z.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 427–434. Association for Computational Linguistics (2005)
7. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics-Volume 2. pp. 539–545. ACL (1992)

8. Heckmann, D., Frank, A., Arnold, M., Gietz, P., Roth, C.: Citation segmentation from sparse & noisy data: An unsupervised joint inference approach with markov logic networks (2013)
9. Jain, D.: Knowledge engineering with Markov Logic Networks: A review. *Evolving Knowledge in Theory and Applications* 16 (2011)
10. Jiang, J., Zhai, C.: A Systematic Exploration of the Feature Space for Relation Extraction. In: *HLT-NAACL*. pp. 113–120 (2007)
11. Kate, R.J., Mooney, R.J.: Joint entity and relation extraction using card-pyramid parsing. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. pp. 203–212. *ACL* (2010)
12. Lafferty, J., McCallum, A., Pereira, F.C.: *Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data* (2001)
13. Li, Q., Ji, H.: Incremental joint extraction of entity mentions and relations. In: *ACL* (2014)
14. Li, Q., Ji, H., Hong, Y., Li, S.: Constructing information networks using one single model. In: *EMNLP* (2014)
15. Lu, W., Roth, D.: Joint mention extraction and classification with mention hypergraphs. In: *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP2015)* (2015)
16. Miwa, M., Bansal, M.: End-to-end Relation Extraction using LSTMs on Sequences and Tree Structures. *arXiv preprint arXiv:1601.00770* (2016)
17. Miwa, M., Sasaki, Y.: Modeling Joint Entity and Relation Extraction with Table Representation. In: *EMNLP*. pp. 1858–1869 (2014)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1), 3–26 (2007)
19. Palshikar, G.K.: *Techniques for Named Entity Recognition. Bioinformatics: Concepts, Methodologies, Tools, and Applications* p. 400 (2013)
20. Qian, L., Zhou, G., Kong, F., Zhu, Q., Qian, P.: Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. pp. 697–704. *ACL* (2008)
21. Richardson, M., Domingos, P.: Markov Logic Networks. *Machine learning* 62(1-2), 107–136 (2006)
22. Roth, D., Yih, W.: A linear programming formulation for global inference in natural language tasks. In: *CoNLL*. pp. 1–8 (2004)
23. Roth, D., Yih, W.t.: Probabilistic reasoning for entity & relation recognition. In: *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. pp. 1–7. *ACL* (2002)
24. Roth, D., Yih, W.t.: Global inference for entity and relation identification via a linear programming formulation. *Introduction to statistical relational learning* pp. 553–580 (2007)
25. Singh, S., Riedel, S., Martin, B., Zheng, J., McCallum, A.: Joint inference of entities, relations, and coreference. In: *Proceedings of the 2013 workshop on Automated knowledge base construction*. pp. 1–6. *ACM* (2013)
26. Singla, P., Domingos, P.: Lifted first-order belief propagation. In: *AAAI*. vol. 8, pp. 1094–1099 (2008)
27. Zhang, C., Hoffmann, R., Weld, D.S.: Ontological Smoothing for Relation Extraction with Minimal Supervision. In: *AAAI* (2012)
28. Zhu, J., Nie, Z., Liu, X., Zhang, B., Wen, J.R.: StatSnowball: a statistical approach to extracting entity relationships. In: *Proceedings of the 18th international conference on World wide web*. pp. 101–110. *ACM* (2009)