Named Entity Recognition: Resource Constrained Maximum Path

Luigi Di Puglia Pugliese[2,⋆], *Elisabetta* Fersini[1,⋆⋆], *Francesca* Guerriero[2,⋆⋆⋆], and *Enza* Messina[1,⋆⋆⋆⋆]

[1] *University of Milano-Bicocca*
[2] *University of Calabria*

Abstract. Information Extraction (IE) is a process focused on automatic extraction of structured information from unstructured text sources. One open research field of IE relates to Named Entity Recognition (NER), aimed at identifying and associating atomic elements in a given text to a predefined category such as names of persons, organizations, locations and so on. This problem can be formalized as the assignment of a finite sequence of semantic labels to a set of interdependent variables associated with text fragments, and can modelled through a stochastic process involving both hidden variables (semantic labels) and observed variables (textual cues). In this work we investigate one of the most promising model for NER based on Conditional Random Fields (CRFs). CRFs are enhanced in a two stages approach to include in the decision process logic rules that can be either extracted from data or defined by domain experts. The problem is defined as a Resource Constrained Maximum Path Problem (RCMPP) associating a resource with each logic rule. Proper resource Extension Functions (REFs) and upper bound on the resource consumptions are defined in order to model the logic rules as knapsack-like constraints. A well-tailored dynamic programming procedure is defined to address the RCMPP.

1 Introduction

Information Extraction (IE) is a task of Natural Language Processing aimed at inferring a structured representation of contents from unstructured textual sources. In this field, Named Entity Recognition (NER) has gained the attention of researches for identifying and associating atomic elements in a given text to a predefined category (such as names of persons, organizations and locations). Considering a text as sequence of tokens $x = x_1, \dots, x_N$, the goal is

[⋆] e-mail: luigi.dipugliapugliese@unical.it
[⋆⋆] e-mail: fersini@disco.unimib.it
[⋆⋆⋆] e-mail: francesca.guerriero@unical.it
[⋆⋆⋆⋆] e-mail: messina@disco.unimib.it

to classify each token x_i as one of the entity labels $y_j \in Y$ for originating a tag sequence $y = y_1, \ldots, y_N$.

Nowadays, the state-of-the-art model for tackling the NER task is represented by linear chain Conditional Random Fields (CRFs) [3], which is a discriminative undirected graphical model able to encode known relationships among tokens (observations) and labels (hidden states). In order to efficiently enhance the description power of CRFs, two main research directions have been investigated to enlarge the information set exploited during training and inference: (1) relaxing the Markov assumption [6] to include long distance dependencies and (2) introducing additional domain knowledge in terms of logical constraints [2, 5]. Considering that the relaxation of the Markov assumption implies an increasing computational complexity, in this paper we focused on the second research direction by formulating the inference task as an Integer Linear Programming problem. In particular, the standard CRFs inference process is enhanced by two main contribution: (1) introducing "extra knowledge" related to semantic constraints about the token labels, and (2) modelling the label assignment problem as a Resource Constrained Maximum Path Problem (RCMPP).

2 Conditional Random Fields

2.1 Background

A CRF [3] is an undirected graphical model that defines a single joint distribution $P(y|x)$ of the predicted labels (hidden states) $y = y_1, ..., y_N$ given the corresponding tokens (observations) $x = x_1, ..., x_N$. Linear Chain CRFs, in which a first-order Markov assumption is made on the hidden variables, define the following conditional distribution:

$$p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \exp\Big(\sum_{t=1}^{N} \sum_{k}^{K} \omega_k f_k(y_t, y_{t-1}x, t)\Big) \tag{1}$$

where $f_k(y_t, y_{t-1}x, t)$ is an arbitrary real-valued feature function over its arguments and ω_k is a learned weight that tunes the importance of each feature function. In particular, when for a token x_t a given feature function f_k is active, the corresponding weight ω_k indicates how to take into account f_k: (1) if $\omega_k > 0$ it increases the probability of the tag sequence y; (2) if $\omega_k < 0$ it decreases the probability of the tag sequence y; (3) if $\omega_k = 0$ has no effect whatsoever. Once the parameters ω_k have been estimated, usually by maximizing the likelihood of the training data [4], the inference phase can be addressed.

2.2 Inference: finding the most probable state sequence

The inference problem in CRF corresponds to find the most likely sequence of hidden state y^*, given the set of observation $x = x_1, ..., x_n$. This problem can be solved by determining y^* such that:

$$y^* = \arg\max_y p(\vec{y}|\vec{x}) \tag{2}$$

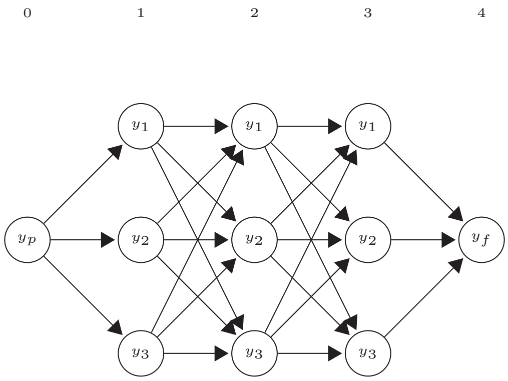


Figure 1. Layered acyclic directed graph.

Given a number m of possible states and n possible input tokens, a layered acyclic directed graph D can be constructed for addressing the inference problem. The graph D is composed of $n+1$ layers. Layer 0 corresponds to the entry layer, $n+1$ is the ending layer, the other n layers represent the elements of the sequence. Arcs from each state y_i, $i = 1, \dots, m$ belonging to each layer t exists for each state y_i, $i = 1, \dots, m$ belonging to layer $t+1$, $t = 0, \dots, n$. We denote as $\mathcal{N}$ the set of nodes containing $2 + n \times m$ elements, that is, the states y_i, $i = 1, \dots, m$ associated with all layers $t = 1, \dots, n$ and two additional states named start state y_p and final state y_f. The set $\mathcal{A}$ contains the arcs (t, y_i, y_i'), $t = 0, \dots, n$, $i = 1, \dots, m$. Arc (t, y_i, y_i') denotes the link of state y_i associated with layer t with the state y_i' associated with layer $t+1$.

A value $\alpha^t_{y_i,y_i'}$ is associated with each arc $(t, y_i, y_i') \in \mathcal{A}$. The parameter $\alpha^t_{y_i,y_i'}$ is proportional to the probability of passing from state y_i at layer t to state y_i' at layer $t+1$. Figure 1 shows the graph D with $n = m = 3$.

Given a layered acyclic directed graph denoting both observed tokens and labels to be assigned, the objective is to find the heaviest path in the graph D starting from y_p and ending at the state y_f. Given the variables $e^t_{y_i,y_i'}$ assuming value equal to 1 if arc (t, y_i, y_i') is included to the optimal path, 0 otherwise, the problem can be formulated as follows.

$$\max \sum_{(t,y_i,y_i') \in \mathcal{A}} \alpha^t_{y_i,y_i'} e^t_{y_i,y_i'} \tag{3}$$

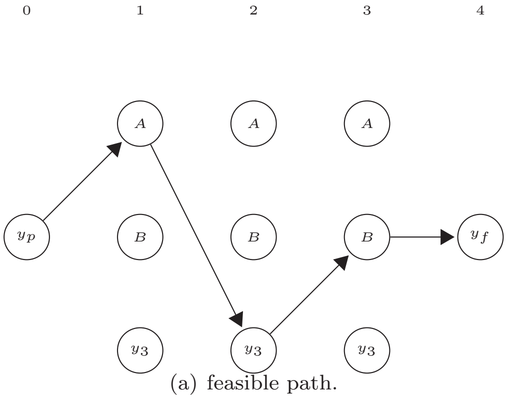
$$s.t.$$

$$\sum_{(t-1,y_i,y_i') \in \mathcal{A}} e^{t-1}_{y_i,y_i'} - \sum_{(t,y_i',y_i) \in \mathcal{A}} e^t_{y_i',y_i} = 0, \ \forall y_i' \in N \setminus \{y_p, y_f\},$$

$$1 \le t \le n \tag{4}$$

$$\sum_{(0,y_p,y_i)} e^0_{y_p,y_i} = 1 \tag{5}$$

$$\sum_{(n,y_i,y_f)} e^n_{y_i,y_f} = 1 \tag{6}$$



(a) feasible path.

Figure 2. Feasible paths for constraints (8).

The following constraints, opportunely instantiated according to the domain knowledge, could be introduced in the model:

- Adjacency: if the token at time $t-1$ is labelled as A, then the token at time t must be labelled as B

$$\sum_{y_i \in \mathcal{N}} e^{t-1}_{y_i,A} - e^{t}_{A,B} \leq 0, \quad t = 1, \ldots, n-1. \tag{7}$$

- Precedence: if the token at time $t+z$ is labelled as B, then a token at time t must be labelled as A

$$\sum_{y_i \in \mathcal{N}} e^{t-1}_{y_i,A} - \sum_{z=1}^{n-t-1} \sum_{y_i \in \mathcal{N}} e^{t+z}_{B,y_i} \geq 0, \quad t = 1, \ldots, n-1. \tag{8}$$

- Begin-end position: if the sequence of tokens starts with label A, then the sequence must end with label B

$$e^{0}_{y_p,A} - e^{n}_{B,y_f} \leq 0. \tag{9}$$

To guarantee constraints (7) it is sufficient to modify the graph (D) by removing all the edges (t, y_i, B). Examples of feasible paths satisfying constraints (8) are depicted in figure 2. Figure 3 shows an infeasible path for the same constraints.

3 Resource constrained model

The problem can be modelled as a Resource Constrained Maximum Path Problem (*RCMPP*). It is possible to define proper Resource Extension Function (*REF*) in order to introduce knapsack-like constraints for each typology of logic rule, that is, precedence and begin-end conditions. The reader is referred to [1] for more details on resource constrained path problems.

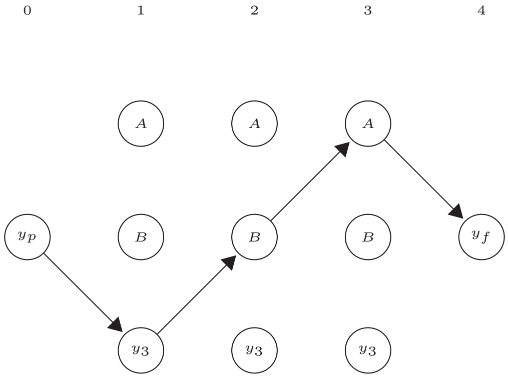


Figure 3. Unfeasible path for constraints (8).

3.1 Precedence Constraint

Let consider P_B be the set of predecessor states of state B. The path from y_p to B has to contain all states $y_{\bar{i}} \in P_B$. We assume a resource consumption $r^p_{i,\bar{i}}$ associated with each state $y_{\bar{i}}$, for each y_i of the network, defined in what follows:

$$r^p_{i,\bar{i}} = \begin{cases} 1, & if \ \ y_i \equiv B; \\ -1, & if \ \ i = \bar{i}; \\ 0, & if \ \ i \neq \bar{i}, y_i \not\equiv B. \end{cases} \tag{10}$$

The *REF* associated with the precedence constraints is defined in Eq. (11).

$$\begin{cases} w^p_{y_p,\bar{i}} = 0, \forall y_{\bar{i}} \in P_B; \\ w^p_{j,\bar{i}} = w^p_{i,\bar{i}} + r^p_{j,\bar{i}}, \forall (t, y_i, y_j) \in \mathcal{A}, y_{\bar{i}} \in P_B \cup \{B\}. \end{cases} \tag{11}$$

The resource limit W_{P_B} is set equal to 0. The set of knapsack-like constraints that define the precedence constraints assume the following form:

$$w^p_{j,\bar{i}} \leq W_{P_B}, \forall j \in \mathcal{N}, \forall y_{\bar{i}} \in P_B. \tag{12}$$

Figure 4 shows the resource constrained instance when precedence constraints are considered.

Considering feasible paths in figure 2, the resource consumption is either 0 or -1 for each label of both paths. The path in Figure 3 has resource consumption equal to 1 at state B, thus it is infeasible for Eq. (12).

It is worth observing that the resource at states can be viewed as resource on arcs. Indeed, the resource consumption of an arc is the resource consumption of the head node.

3.2 Begin-end Constraint

Here we define resource constraint for the begin-and condition. We assume a resource consumption $r^{be}_{(t,y_i,y_j)}$ associated with each arc $(t, y_i, y_j) \in \mathcal{A}$.

DOI: 10.1051/itmconf/20171400004

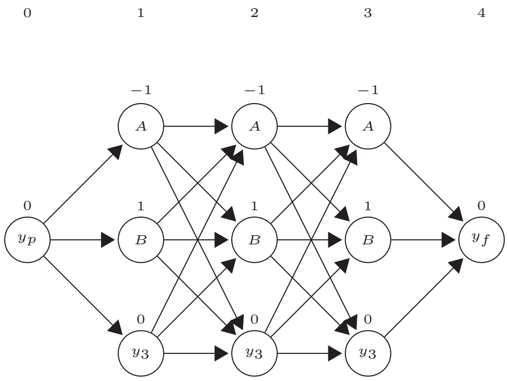


Figure 4. Resource constrained graph with the precedence constraint $P_B = \{A\}$.

$$r^{be}_{(t,y_i,y_j)} = \begin{cases} 1, & if \ \ t = 0, y_i = y_p, y_j = A; \\ -1, & if \ \ t = n, y_i = B, y_j = y_f; \\ 0, & for \ all \ other \ cases. \end{cases} \tag{13}$$

The REF associated with begin-and constraint is defined in equation (14).

$$\begin{cases} w^{be}_{y_p} = 0; \\ w^{be}_j = w^{be}_i + r^{be}_{(t,y_i,y_j)}, \forall (t, y_i, y_j) \in \mathcal{A}. \end{cases} \tag{14}$$

Assuming a resource limit $W_{be} = 0$, the resource constraint modelling begin-end condition is expressed by equation (15).

$$\sum_{(t,y_i,y_i') \in \mathcal{A}} e^t_{y_i,y_i'} r^{be}_{(t,y_i,y_i')} \leq W_{be}. \tag{15}$$

Figure 5 shows the resource graph associated with the begin-end condition.

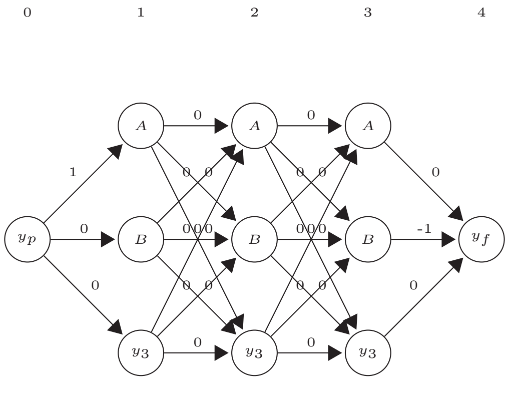


Figure 5. Resource constrained graph with begin-end condition.

4 Sketch of dynamic programming

Let l_i^t be a label associated with state y_i at layer t. The label l_i^t maintains information related to the objective function (α_i^t) and to the resource consumption ($w_{i,\bar{i}}^p, w_i^{be}$). A path π_i^t is associated with each label $l_i^t(\alpha_i^t, w_{i,\bar{i}}^p, w_i^{be})$. Since there may exist several paths to state y_i, the index h is used to indicate the id of a given path. Thus, label $l_i^t(h)$ is associated to the $h-th$ path $\pi_i^t(h)$. Starting from the initial label $l_{y_p}^0(0,0,0)$, the dynamic programming explores the state-space in order to reach the final labels $l_{y_f}^n(h)(\cdot)$. Among all labels $l_{y_f}^n(h)(\cdot)$, that with maximum value of α is associated with optimal path. The state-space is reduced by considering only feasible and non-dominated labels.

Definition 4.1 (Feasibility)**.** A label $l_i^t(h)$ is feasible if and only if $w_{i,\bar{i}}^p(h) \leq W_{P_B}, \forall y_{\bar{i}} \in P_B$ and $w_i^{be}(h) \leq W_{be}$, with $i \equiv y_f$.

Definition 4.2 (Dominance)**.** Given two labels $l_i^t(h)$ and $l_i^t(h')$, the first dominates the second if the following conditions hold

$$\alpha_i^t(h) \geq \alpha_i^t(h'); \tag{16}$$

$$w_{i,\bar{i}}^p(h) \leq w_{i,\bar{i}}^p(h'); \tag{17}$$

$$w_i^{be}(h) \leq w_i^{be}(h'); \tag{18}$$

and at least one inequality is strictly satisfied.

Let L be the list of labels with the potential to generate an optimal solution and ND_i be the set of non-dominated labels associated with y_i. The labelling approach to solve to optimality the problem is depicted in Algorithm 1.

Algorithm 1 . $\mathcal{DP}$ scheme

```
Step 0 (Initialization Phase)
Set: $L = \{l_{y_p}^0(0)\}$.

Step 1 (Label Selection)
Select and delete from $L$ a label $l_i^t(h)$.

Step 2 (Labels Generation)
for all $y_j : (t, y_i, y_j) \in \mathcal{A}$ do
    Compute $\alpha_j^{t+1}$, $w_{j,\bar{i}}^p, \forall y_{\bar{i}} \in P_B$, and $w_j^{be}$.
    if Label $l_j^{t+1}(\alpha_j^{t+1}, w_{j,\bar{i}}^p, w_j^{be})$ is feasible then
        if Label $l_j^{t+1}$ is not dominated by any labels in $ND_j$ then
            Add label $l_j^{t+1}$ to $ND_j$ and $L$.
            Remove from $ND_j$ all labels dominated by $l_j^{t+1}$.
        end if
    end if
end for

Step 3 (Termination check)
if $L = \emptyset$ then
    STOP
else
    Go to Step 1.
end if
```

5 Conclusion

In this paper, the problem of Named Entity Recognition is addressed by investigating the inference task on Conditional Random Fields. In particular, a mathematical programming formulation based on a Resource Constrained Maximum Path is presented to include some background knowledge during the labelling phase of a text source. Three types of background knowledge constraints have been presented, together with a dynamic programming approach for determining the optimal sequence of labels. Concerning the future work, additional long distance dependencies are planned to be automatically discovered from the data (as hidden patterns) and enclosed into the mathematical problem formulation.

References

[1] L. Di Puglia Pugliese and F. Guerriero. A survey of resource constrained shortest path problems: Exact solution approaches. *Networks*, 62(3):183–200, 2013.

[2] E. Fersini, E. Messina, G. Felici, and D. Roth. Soft-constrained inference for named entity recognition. *Information Processing & Management*, 50(5):807–819, 2014.

[3] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th Int. Conf on Machine Learning*, 2001.

[4] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proc. of the 6th Conf. on Natural language learning*, 2002.

[5] D. Roth and W.-T. Yih. Ilp inference for conditional random fields. In *Proc. of the 22nd Int. Conf. on ML*, 2005.

[6] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *Proc. of the 18th Conf. on Neural Information Processing Systems*, 2004.