Information Processing and Management 50 (2014) 807–819

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Soft-constrained inference for Named Entity Recognition

E. Fersini [a,*], E. Messina [a], G. Felici [b], D. Roth [c]

[a] DISCo, University of Milano – Bicocca, Viale Sarca, 336, 20126 Milano, Italy
[b] CNR, Institute for Systems Analysis and Computer Science, Viale Manzoni, 30, 00185 Roma, Italy
[c] Department of Computer Science, University of Illinois at Urbana–Champaign, 2700 Prairie Meadow Dr., Champaign, IL 61822, United States

ARTICLE INFO

Article history:
Received 30 April 2013
Received in revised form 14 April 2014
Accepted 22 April 2014
Available online 17 May 2014

Keywords:
Conditional Random Fields
Named Entity Recognition
Rule extraction
Integer linear programming

ABSTRACT

Much of the valuable information in supporting decision making processes originates in text-based documents. Although these documents can be effectively searched and ranked by modern search engines, actionable knowledge need to be extracted and transformed in a structured form before being used in a decision process. In this paper we describe how the discovery of semantic information embedded in natural language documents can be viewed as an optimization problem aimed at assigning a sequence of labels (hidden states) to a set of interdependent variables (textual tokens). Dependencies among variables are efficiently modeled through Conditional Random Fields, an indirected graphical model able to represent the distribution of labels given a set of observations. The Markov property of these models prevent them to take into account long-range dependencies among variables, which are indeed relevant in Natural Language Processing. In order to overcome this limitation we propose an inference method based on Integer Programming formulation of the problem, where long distance dependencies are included through non-deterministic soft constraints.

1. Introduction

The data used by Decision Support Systems are assumed to be structured and quantifiable. However, thanks to the growing of the web and the spread of Document Management Systems, most of the valuable information are embedded in textual documents that need to be processed to extract relevant information in a machine readable form to become actionable. In most of the cases this activity involves the analysis of human language texts by means of Natural Language Processing (NLP) techniques. Named Entity Recognition (NER) is the task aimed at identifying and associating atomic elements in a given text to a set of predefined categories such as names of persons, organizations, locations, dates, and quantities.

Early NER systems have been defined as rule-based approaches with a set of fixed and manually coded rules provided by domain experts (Rau, 1991; Lehnert et al., 1993; Riloff, 1993; Appelt, Hobbs, Bear, Israel, & Tyson, 1993). Considering the costs, in terms of human effort, to reveal and formulate hand-crafted rules, several research communities ranging from Statistical Analysis to Natural Language Processing and Machine Learning have provided valuable contributions for automatically derive models able to detect and categorize pre-defined entities. The first tentatives, aimed at deriving these rules under the form of boolean conditions, are based on inductive learner where rules can be learnt automatically from labeled examples. The inductive rule learning approach has been instantiated according to different learning paradigm: bottom-up

[*] Corresponding author. Tel.: +39 0264487919; fax: +39 0264487880.
E-mail addresses: fersini@disco.unimib.it (E. Fersini), messina@disco.unimib.it (E. Messina), giovanni.felici@iasi.cnr.it (G. Felici), danr@Illinois.edu (D. Roth).

http://dx.doi.org/10.1016/j.ipm.2014.04.005
0306-4573/© 2014 Elsevier Ltd. All rights reserved.

(Califf & Mooney, 2003; Califf & Mooney, 1999; Ciravegna, 2001), top-down (Soderland, 1999; Quinlan, 1990; Landwehr, Kersting, & Raedt, 2007; Ho & Nguyen, 2003) and interactive rule learning (Khaitan, Ramakrishnan, Joshi, & Chalamalla, 2008; Bohannon et al., 2009; Beckerle, Martucci, & Ries, 2010).

An alternative approach to inductive rule learners is represented by statistical methods, where the NER task is viewed as a decision making process aimed at assigning a sequence of labels to a set of either joint or interdependent variables, where also complex relationships may hold among them. This decision making paradigm can be addressed in two different ways: (1) at segment-level (Sarawagi & Cohen, 2004; Daumé & Marcu, 2005; Galen, 2006), where the NER task is managed as a segmentation problem in which each segment corresponds to an entity label and (2) at token level (Takeuchi & Collier, 2002; Seymore, McCallum, & Rosenfeld, 1999; Ratnaparkhi, 1999; Richardson & Domingos, 2006; Lafferty, McCallum, & Pereira, 2001), where an entity label is assigned to each token of the sentence. In the first case the output of the decision process is a sequence of segments. More formally, a segmentation s of an input sentence $x = x_1, \ldots, x_N$ is a sequence of segments $s_1 \ldots s_p$ with $p \leqslant N$. Each segment s_j consists of a start position l_j, an end position u_j, and a label y belonging to a set of entity labels $\mathcal{Y}$. The second decision making paradigm is represented by token-level models, where the unstructured text is tackled as a sequence of tokens and the output of the decision process is a sequence of labels $y = y_1, \ldots, y_N$.

Nowadays, the state of the art to model a NER problem is represented by Linear Chain Conditional Random Fields (Lafferty et al., 2001). This model, thanks to its advantages over generative approaches, has been extensively investigated to extract named entities from different unstructured sources such as judicial transcriptions (Fersini, Messina, Archetti, & Cislaghi, 2013; Fersini & Messina, 2013), medical reports (Cvejic, Zhang, Marx, & Tjoe, 2012; Deléger et al., 2013), and user generated contents (Qi & Chen, 2010; Shariaty & Moghaddam, 2011). The efficiency of Linear Chain Conditional Random Fields (CRF) is strictly related to the underlying Markov assumption: given the observation of a token, the corresponding hidden state (label) depends only on the labels of its adjacent tokens. In order to efficiently enhance the description power of CRF, during the last ten years several approaches have been proposed to enlarge the information set exploited during training and inference. In particular, two main research directions have been investigated: (1) relaxing the Markov assumption (Sarawagi & Cohen, 2004; Galen, 2006) to model long distance relationships and (2) introducing additional domain knowledge in terms of logical constraints during the inference phase (Kristjansson, Culotta, Viola, & McCallum, 2004; Roth & Yih, 2005; Chang, Ratinov, & Roth, 2007; Chang, Ratinov, & Roth, 2008). Considering that the relaxation of the Markov assumption implies an increasing computational complexity related to the training and inference phase, in this paper we focused our attention on an integer linear programming (ILP) formulation of the inference problem by including soft constraints obtained by learning declarative rules from data.

The introduction of constraints allows us to improve the global label assignment by correcting mistakes of local predictions. The label assignment problem is therefore solved through a constrained optimization problem where the extra-knowledge related to complex relationships among variables is represented through a set of logical rules easily introduced as linear inequalities. This approach, as shown by the experimental results, makes it possible to significantly improve the performances of CRF in NER tasks.

The outline of the paper is the following. In Section 2 a brief review of CRF is presented along with a background overview of the training phase, together with the most relevant inference approaches able to include domain constraints. In Section 3 the proposed soft-constrained inference approach is detailed by focusing on learning constraints from data and by presenting its mathematical programming formulation. In Section 4 the experimental investigation on both benchmark and datasets is described, while in Section 5 conclusions and ongoing research are summarized.

2. Conditional Random Fields

A Conditional Random Field is an indirected graphical model that defines the joint distribution $P(y|x)$ of the predicted labels (hidden states) $y = y_1, \ldots, y_N$ given the corresponding tokens (observations) $x = x_1, \ldots, x_N$. Now, consider X as the random variable over data sequences (natural language sentences) to be labeled, and Y is the random variable over corresponding label sequences over a finite label alphabet $\mathcal{Y}$. The joint distribution $P(X, Y)$ is represented by a conditional model $P(Y|X)$ from paired observation and label sequences, and the marginal probability $p(X)$ is not explicitly model. The formal definition of CRF (Lafferty et al., 2001) is given below:

Definition 1 (*Conditional Random Fields*). Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a Conditional Random Field, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

Thus, a CRF is a random field globally conditioned on the observation X. Throughout the paper we tacitly assume that the graph G is fixed. A Linear-Chain Conditional Random Field is a Conditional Random Field in which the output nodes are linked by edges in a linear chain. The graphical representation of a general CRF and a Linear-Chain CRF is reported in Fig. 1. In the following, Linear-Chain CRF are assumed.

According to the Hammersley–Clifford theorem (Hammersley & Clifford, 1971; Clifford, 1990), given $\mathcal{C}$ as the set of all cliques in G, the conditional probability distribution of a sequence of labels y given a sentence x can be written as:

$$p(y|x) = \frac{1}{Z(x)} \prod_{C \in \mathcal{C}} \Phi_C(x_C, y_C) \tag{1}$$