# Beyond Linear Chain: A Journey through Conditional Random Fields for Information Extraction from Text

Diego Marcheggiani
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy
*diego.marcheggiani@isti.cnr.it*

**Abstract**

Information Extraction (IE) is a field at the crossroads of IR and NLP that studies methods for extracting information from text in such a way that this information can be used to populate a structured information repository. The main methods by means of which IE has been tackled rely on supervised learning;the best-performing such methods belong to the class of *probabilistic graphical models*, and, in particular, to the class of *Conditional Random Fields* (CRFs). In this thesis we investigate two major aspects related to textual IE via CRFs: (a) the creation of CRFs models that can outperform the commonly adopted linear-chain CRFs, and the creation of methods for ensuring the quality of training data and for assessing the impact of training data quality on the accuracy of CRFs systems for IE.

We start by facing the task of IE from medical documents written in the Italian language. We propose two novel approaches: (i) a cascaded, two-stage method composed by two layers of CRFs, and (ii) a confidence-weighted ensemble method that combines standard linear-chain CRFs and the proposed two-stage method. Both the proposed models are shown to outperform a standard linear-chain CRFs system.

We then investigate aspect-oriented sentence-level opinion mining from product reviews, that consists in predicting, for all sentences in the review, whether the sentence expresses a positive, neutral, or negative opinion (or no opinion at all) about a specific aspect of the product. We propose a set of increasingly powerful models based on CRFs, including a hierarchical multi-label CRFs scheme that jointly models the overall opinion expressed in a product review and the set of aspect-specific opinions expressed in each of its sentences. The proposed CRFs models are shown to obtain better results than linear-chain CRFs.

We then study the impact that the quality of training data has on the accuracy of an IE system via experiments performed on a dataset in which inter-coder agreement data are available. Finally, we investigate active learning techniques for a type of semi-supervised CRFs specifically devised for partially labeled sequences. We show that margin-based strategies always obtain the best results on the four tasks we have tested them on.

Soon available at `http://nmis.isti.cnr.it/marcheggiani`.

**Supervisors**: Andrea Esuli (ISTI-CNR), Fabrizio Sebastiani (ISTI-CNR).