A Metaheuristic for Named Entity Recognition

D. Ferone, E. Fersini, E. Messina

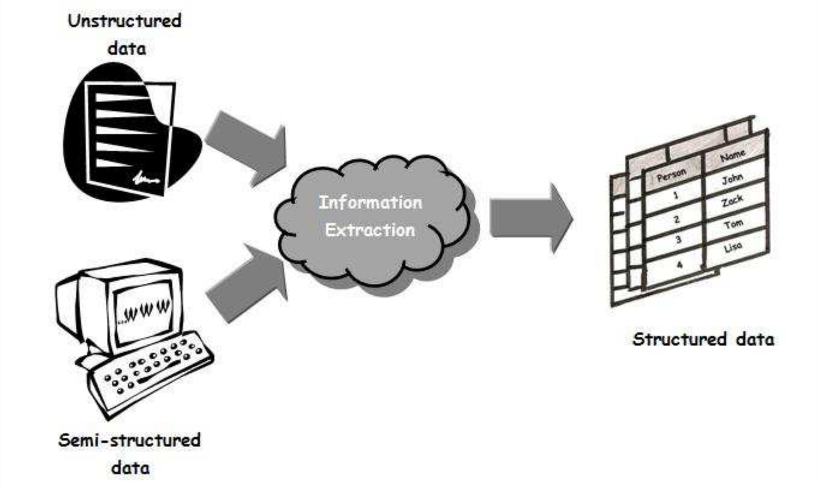
Department of Informatics, Systems and Communication, University of Milano-Bicocca

Table of contents

Introduction & CRF

Information Extraction

Information Extraction (IE) aims at extracting structured information from non-structured or semi-structured texts



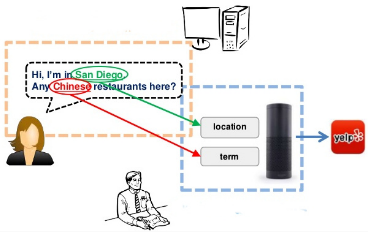
Name Entity Recognition

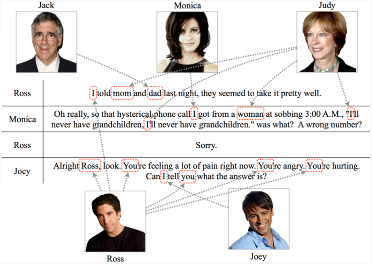
Named Entity Recognition (NER) is an IE task that seeks to locate and classify text segments into predefined classes/labels (e.g., Person, Location, Organization)

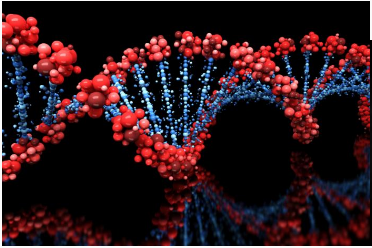
CRICKET - MILLNS SIGNS FOR BOLAND
CAPE TOWN 1996-08-22
South African provincial side Boland said on Thursday they had signed Leicestershire fast bowler David Millns on a one year contract. Millns, who toured Australia with England A in 1992, replaces former England all-rounder Phillip DeFreitas as Boland's overseas professional.

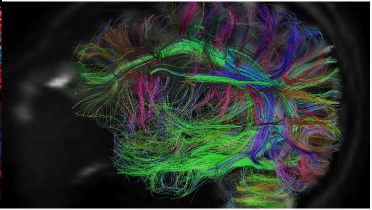
Labels	Examples
PER	David Millns Philip DeFreitas
ORG	Boland
LOC	Cape Town England Australia

Applications



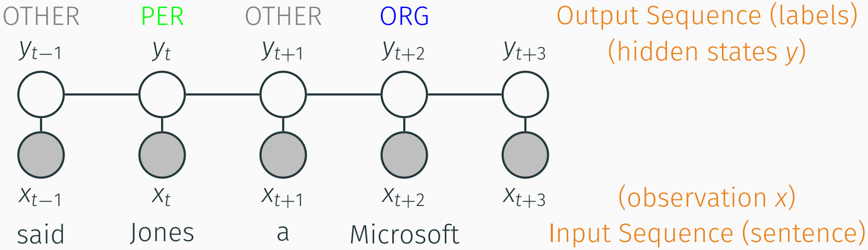






Conditional Random Fields

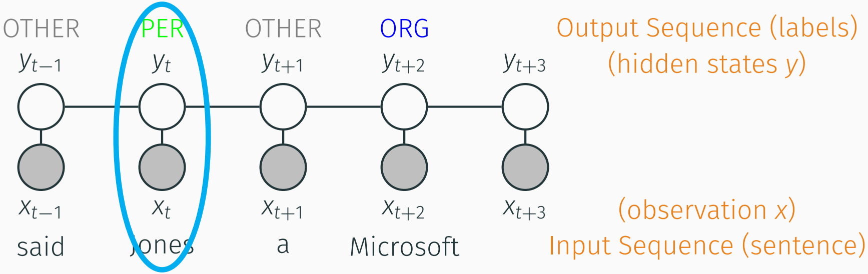
Consider X as the random variable over data sequences (natural language sentences) to be labeled, and Y is the random variable over corresponding label sequences over a finite label alphabet Y.



$$P(y|x) = \frac{\exp \sum_{t=1}^{T} \left(\sum_i \lambda_i f_i(y_t, x) + \sum_j \mu_j g_j(y_t, y_{t-1}, x) \right)}{Z(x)}$$

Conditional Random Fields

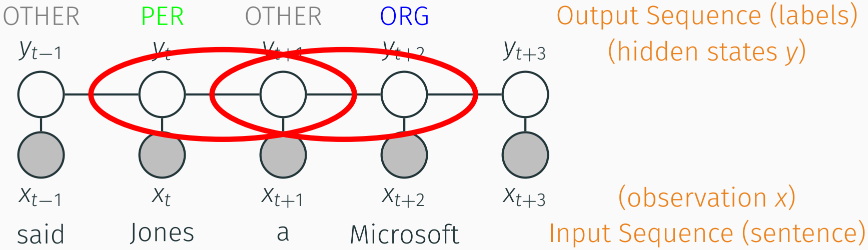
Consider X as the random variable over data sequences (natural language sentences) to be labeled, and Y is the random variable over corresponding label sequences over a finite label alphabet Y.



$$P(y|x) = \frac{\exp \sum_{t=1}^{T} \Big(\sum_i \lambda_i f_i(y_t, x) + \sum_j \mu_j g_j(y_t, y_{t-1}, x) \Big)}{Z(x)}$$

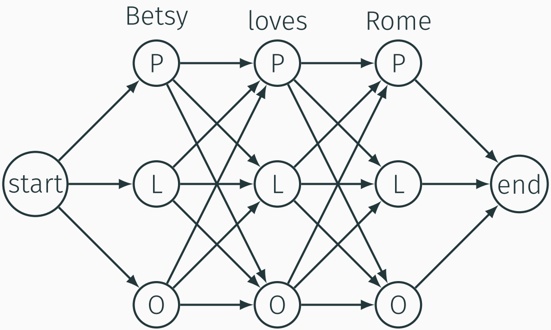
Conditional Random Fields

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 Y.



$$P(y|x) = \frac{\exp \sum_{t=1}^{T} \Big(\sum_i \lambda_i f_i(y_t, x) + \sum_j \mu_j g_j(y_t, y_{t-1}, x) \Big)}{Z(x)}$$

Layered graph



Definition: 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.

Mathematical model

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

$$y = \arg\max p(y|x).$$

Mathematical model

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

$$y = \arg\max p(y|x).$$

$$\max \sum_{\psi_{ii'}^t \in A} e_{ii'}^t \alpha_{ii'}^t \tag{1a}$$

$$\text{s.t.} \sum_{i_1=0}^{m-1} e_{i_1 i}^{t-1} - \sum_{i_2=0}^{m-1} e_{i i_2}^t = 0, \qquad \forall i \in V \setminus \{start, end\}, 1 \leq t \leq n \tag{1b}$$

$$\sum_{\psi_{start,i}^0 \in A} e_{start,i}^0 = 1 \tag{1c}$$

$$\sum_{\psi_{i,end}^n \in A} e_{i,end}^n = 1 \tag{1d}$$

$$e_{ij}^t \in \{0, 1\} \qquad \forall i, j, t \text{ s.t. } 0 \leq i, j < m, 0 \leq t < n. \tag{1e}$$

Constrained CRF

Complex relationships

- CRFs are very good in capturing capture local properties;
- Very efficient thanks to the Markovian assumption and Viterbi algorithm;
- but not able to model complex relationships;
- add additional constraints to CRF to model them.

Constraints examples

Adjacency

$$\sum_{i_1=0}^{m-1} e_{i_1 A}^{t-1} - \sum_{i_2=0}^{m-1} e_{B i_2}^{t} \leq 0$$

Precedence

$$\sum_{i_1=0}^{m-1} e_{i_1 A}^{t-1} - \sum_{z=0}^{n-t} \sum_{i_2=0}^{m-1} e_{B i_2}^{t+z} \leq 0$$

State change

$$\sum_{i_1=0}^{m-1} 2(e_{i_1 d}^{t-1}) - \sum_{i_2=0}^{m-1} e_{i_2 A}^{t-2} - \sum_{i_3=0}^{m-1} e_{B i_3}^{t} \leq 0$$

Constraints examples

Begin-end

$$\sum_{i_1=0}^{m-1} e^1_{Ai_1} - \sum_{i_2=0}^{m-1} e^{n-1}_{i_2B} \leq 0$$

Presence and Precedence

$$m(t-2)e^t_{Ai_1} - \sum_{z=1}^{t-2}\sum_{i_2=0}^{m-1}(1 - e^z_{i_2B}) \leq 0, \text{ with } 2 \leq t \leq n \text{ and } 0 \leq i_1 \leq m-1$$

Feasibily problem

- These rules are not necessarily satisfied by all training sentences and by data during the inference phase;
- Introducing these constraints as hard-constraints could lead to feasibility problems.

Feasibily problem

- These rules are not necessarily satisfied by all training sentences and by data during the inference phase;
- Introducing these constraints as hard-constraints could lead to feasibility problems.

$$\Downarrow$$

Two step approach:

Feasibily problem

- These rules are not necessarily satisfied by all training sentences and by data during the inference phase;
- Introducing these constraints as hard-constraints could lead to feasibility problems.

$$\Downarrow$$

Two step approach:

1. solve the model without constraints;

Feasibily problem

- These rules are not necessarily satisfied by all training sentences and by data during the inference phase;
- Introducing these constraints as hard-constraints could lead to feasibility problems.

$$\Downarrow$$

Two step approach:

1. solve the model without constraints;
2. introduce constraints that can be violated (soft-constraints), but with a penalization cost. Minimize the cost of violating the constraints approximating the solution of the first step.

Resulting problem

$$
\begin{aligned}
&\min && && \text{(2a)}\\
&\text{s.t.} \sum_{i_1=0}^{m-1} e_{i_1 i}^{t-1} - \sum_{i_2=0}^{m-1} e_{i i_2}^{t} = 0, && \forall i \in V \setminus \{start, end\}, 1 \le t \le n && \text{(2b)}\\
&\sum_{\psi_{start,i}^{0} \in A} e_{start,i}^{0} = 1 && && \text{(2c)}\\
&\sum_{\psi_{i,end}^{n} \in A} e_{i,end}^{n} = 1 && && \text{(2d)}\\
&e_{ij}^{t} \in \{\, 0, 1 \,\} && \forall i, j, t \text{ s.t. } 0 \le i, j < m, 0 \le t < n. && \text{(2e)}
\end{aligned}
$$

Resulting problem

$$\min \sum_{h} c_h \sigma_h \tag{2a}$$

$$\text{s.t.} \sum_{i_1=0}^{m-1} e_{i_1 i}^{t-1} - \sum_{i_2=0}^{m-1} e_{i i_2}^{t} = 0, \qquad \forall i \in V \setminus \{start, end\}, 1 \leq t \leq n \tag{2b}$$

$$\sum_{\psi_{start,i}^{0} \in A} e_{start,i}^{0} = 1 \tag{2c}$$

$$\sum_{\psi_{i,end}^{n} \in A} e_{i,end}^{n} = 1 \tag{2d}$$

$$e_{ij}^{t} \in \{\, 0, 1 \,\} \qquad \forall i, j, t \text{ s.t. } 0 \leq i, j < m, 0 \leq t < n. \tag{2e}$$

$$\sum_{\psi_{ii'}^{t} \in A} e_{ii'}^{t} \alpha_{ii'}^{t} \geq \tau \cdot V \tag{2f}$$

$$L \cdot e - \sigma \leq 0 \tag{2g}$$

Solution approach

Iterated Local Search

- Meta-heuristic framework that iteratively applies local search, perturbation, and evaluation of the solution against an acceptance criterion;
- local search performs the intensification phase;
- the perturbation and the acceptance criterion allow to explore the search space as well as to escape from local optima (diversification phase).

Iterated Local Search

```
Function ILS(inputs, parameters)
    baseSol ← createInitialSolutions(inputs, parameters)
    betsSol ← baseSol
    while stopping criterion not reached do
        newSol ← shake(baseSol)
        improving ← True
        while improving do
            newSol ← local-search(newSol)
            if cost(newSol) < cost(baseSol) then
                baseSol ← newSol
                if cost(newSol) < cost(bestSol) then
                    bestSol ← newSol
            else
                improving ← False
                if acceptance-criterion(baseSol) then
                    baseSol ← newSol
    return bestSol
```

Implementation details

Construction Viterbi algorithm (maximum path)

Implementation details

Construction Viterbi algorithm (maximum path)

Shaking re-assign a p% of tokens to random categories

Implementation details

Construction Viterbi algorithm (maximum path)

Shaking re-assign a p% of tokens to random categories

Acceptance criterion if $cost(newSol) < cost(baseSol)$ then accept, otherwise accent *newSol* with a given probability $0 \leq s \leq 1$

Local searches: Adjust punctuation

Correct state changes in non-punctuation tokens.

"M. Kitsuregawa, H. Tanaka, and T. Moto-oka. Application of hash to data base machine and its architecture. New Generation Computing, 1(1), 1983."

Local searches: Adjust punctuation

Correct state changes in non-punctuation tokens.

"M. Kitsuregawa, H. Tanaka, and T. Moto-oka. Application of hash to data base machine and its architecture. New Generation Computing, 1(1), 1983."

Title: Application of hash to
BookTitle: data base machine and its architecture.

Local searches: Adjust punctuation

Correct state changes in non-punctuation tokens.

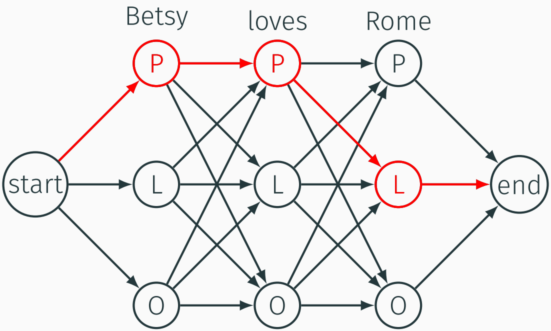
"M. Kitsuregawa, H. Tanaka, and T. Moto-oka. Application of hash to data base machine and its architecture. New Generation Computing, 1(1), 1983."

Title: Application of hash to
BookTitle: data base machine and its architecture.

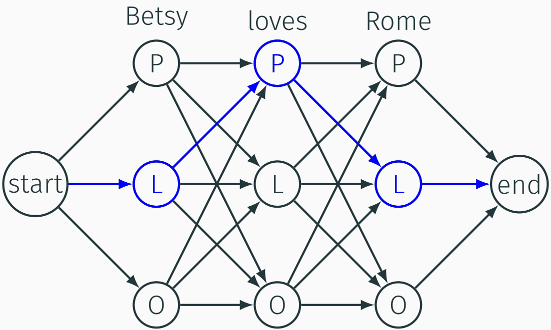
$\Downarrow$

Title: Application of hash to data base machine and its architecture.

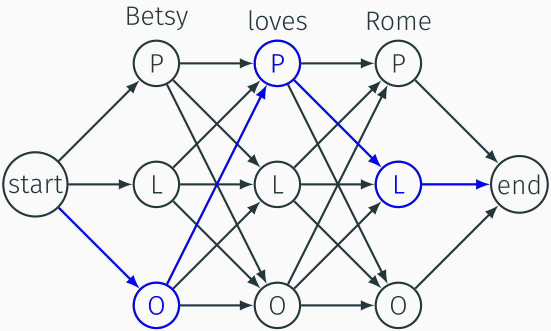
Local searches: general local search



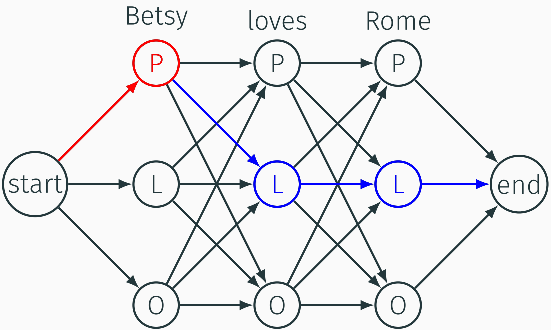
Local searches: general local search



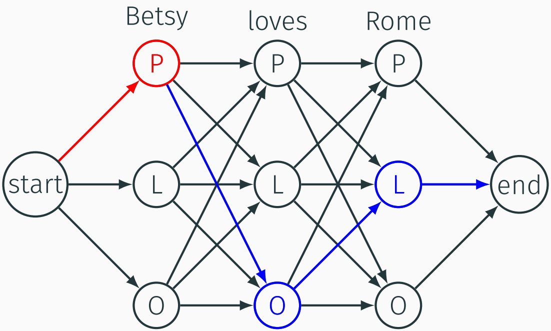
Local searches: general local search



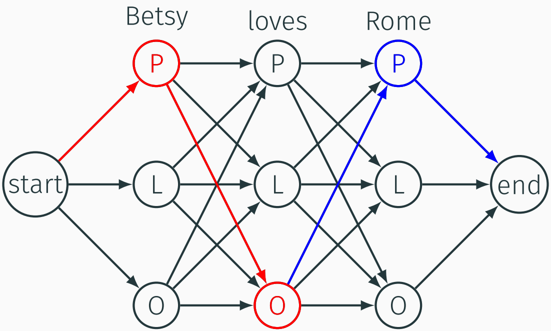
Local searches: general local search



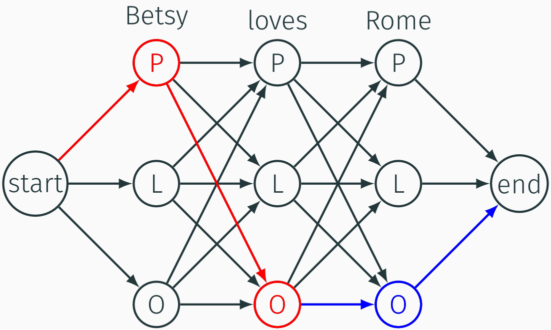
Local searches: general local search



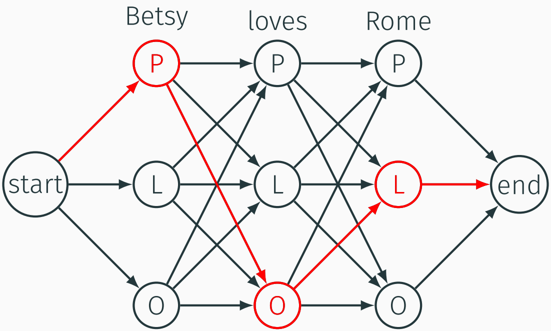
Local searches: general local search



Local searches: general local search



Local searches: general local search



Results and conclusions

Testing settings

Computer: Linux Ubuntu 18.04; Intel® Core™ i7-4510U CPU @ 2.00GHz × 4; 8GB RAM.

Dataset: *Cora* citation benchmark composed of 500 citations of research papers annotated with 13 different labels: *Title, Author, Publisher, Book Title, Date, Journal, Volume, Tech, Institution, Pages, Editor, Location, Notes*.

Testing settings

Constraints:

Start	The citation can only start with author or editor.
AppearsOnce	Each field must be a consecutive list of words, and can appear at most once in a citation.
Punctuation	State transitions must occur on punctuation marks.
BookJournal	The words proc, journal, proceedings, ACM are JOURNAL or BOOKTITLE.
Date	Four digits starting with 20xx and 19xx are DATE.
Editors	The words ed, editors correspond to EDITOR.
Journal	The word journal is JOURNAL.
Note	The words note, submitted, appear are NOTE.
Pages	The words pp., pages correspond to PAGE.
TechReport	The words tech, technical are TECH_REPORT.
Title	Quotations can appear only in titles.
Location	The words CA, Australia, NY are LOCATION.

Results

	ILS	Cplex	HMM^{CCM}
Average F-score	0.77	0.85	
Weighted Average F-score	0.87	0.90	
Accuracy	0.88	0.91	0.94
Time	1.41s	69.49s	

Conclusions and future work

- Constraints can be used only during inference;
- improve the local search phase: **segments**;
- other datasets: US50, CoNLL-2003, Advertisements.

01000101 01001110 01000100

(E N D)

Thank you.