Named Entity Extraction using AdaBoost

Xavier Carreras, Lluís Màrquez & Lluís Padró

TALP Research Center

UPC
Technical University of Catalonia

CoNLL'02 Shared Task

Approach

- Named Entity Extraction as two separate modules:

$$\boxed{\textbf{NER}} \rightarrow \boxed{\textbf{NEC}}$$

- Learning Algorithm: AdaBoost
  - binary classification, binary features
  - real-valued AdaBoost with confidence-rated predictions.

    [Schapire & Singer 99]
  - $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$
  - Weak Rules (h_t): Decision Trees of fixed depth.
  - Good Performance in NLP domains.

Features

- Sliding Window, codifying ± 3 words.
- Primitive Features in the window:
  - Word form
  - PoS (when available)
  - Ortographic Features

initial-caps	*all-caps*	*all-digits*
roman-number	*contains-dots*	*contains-hyphen*
acronym	*lonely-initial*	*punctuation-mark*
single-char	*functional-word*	*URL*

  - Word-Type Patterns:
    functional uppercase lowercase punctuation quote other
  - Bag-of-Words (± 5 window)
  - Trigger Words (class of); also patterns
  - Gazetteer Features (class of)
  - Left Predictions

NER Decision Schemes

- BIO:

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
B	-1	2	1	1	2	4	1	-1
I			2	-1		2	2	1
O	1	-1	-1	2	-1	-1	-1	3
output	O	B	I	O	B	B	I	O

- greedy Open-Close, plus Inside checking:

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Open	-1	1		-1	2	3		-1
I			3				2	-1
Close		-1	2		3	-2	-1	
output	O	B	I	O	B	B	I	O

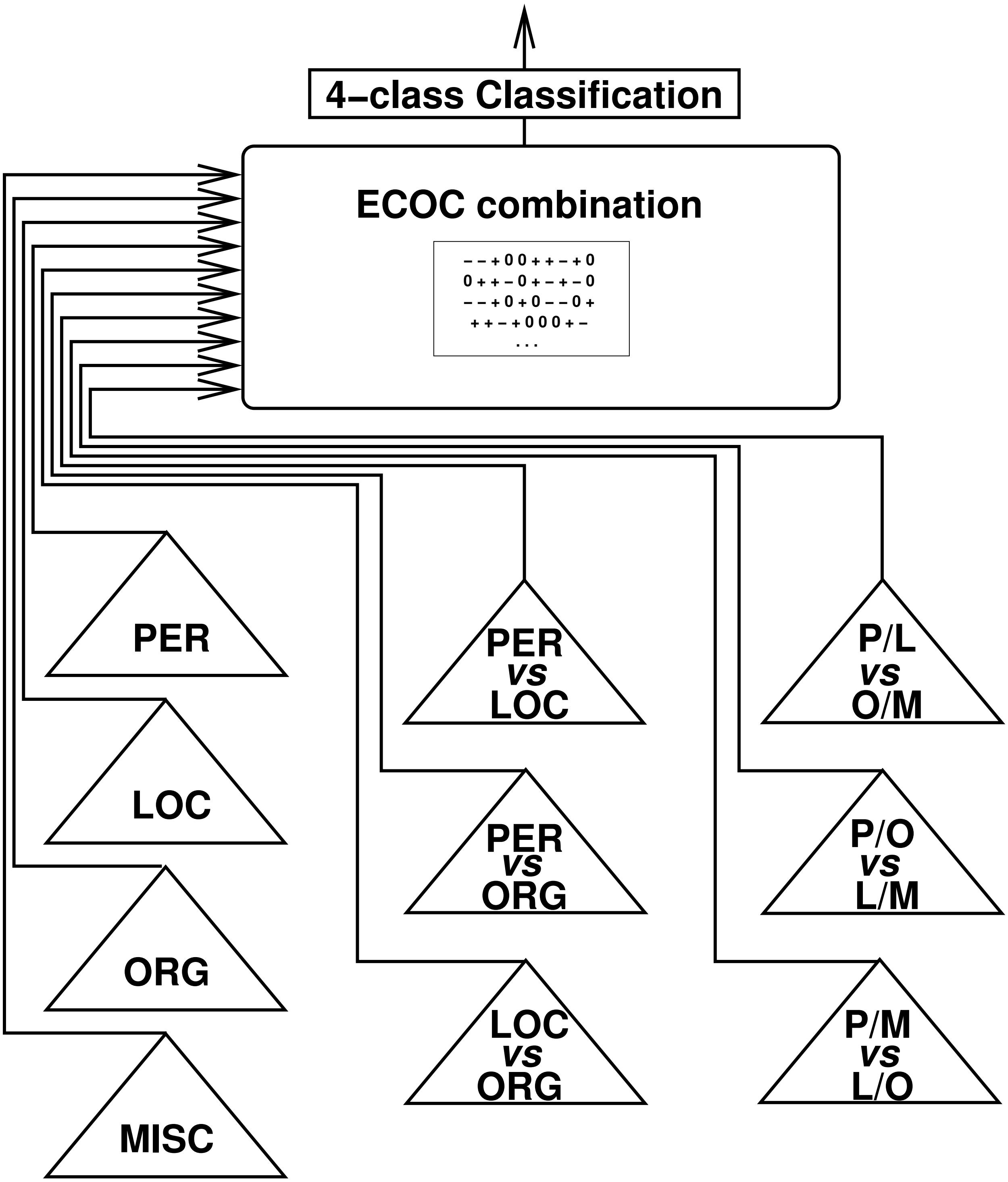
- global Open-Close:

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Open	x	O	x	x	O	O	x	x
Close	x	x	C	x	x	x	C	x
gl 1	O	B	I	I	I	I	I	O
gl 2	x	B	I	O	B	I	I	O
gl 3	x	B	I	O	B	B	I	O
output	O	B	I	O	B	B	I	O

NER Decision Schemes

Spanish dev.	prec.	recall	F_1
BIO	92.45	90.88	91.66
greedy OC	90.50	90.40	90.45
global OC	91.58	90.21	90.89

NEC Module



4–class Classification
ECOC combination
− − + 0 0 + + − + 0
0 + + − 0 + − + − 0
− − + 0 + 0 − − 0 +
+ + − + 0 0 0 + −
. . .
PER
LOC
ORG
MISC
PER vs LOC
PER vs ORG
LOC vs ORG
P/L vs O/M
P/O vs L/M
P/M vs L/O

Results on Spanish

Spanish dev.	precision	recall	$F_{\beta=1}$
LOC	79.04%	80.00%	79.52%
MISC	55.48%	54.61%	55.04%
ORG	79.57%	76.06%	77.77%
PER	87.19%	86.91%	87.05%
overall	79.15%	77.80%	78.47%
overall, no resources	77.91%	76.59%	77.24%

Spanish test	precision	recall	$F_{\beta=1}$
LOC	85.76%	79.43%	82.47%
MISC	60.19%	57.35%	58.73%
ORG	81.21%	82.43%	81.81%
PER	84.71%	93.47%	88.87%
overall	81.38%	81.40%	81.39%
overall, no resources	79.27%	79.29%	79.28%

Results on Dutch

Dutch devel.	precision	recall	$F_{\beta=1}$
LOC	69.71%	80.25%	74.61%
MISC	80.50%	73.59%	76.89%
ORG	76.79%	69.66%	73.05%
PER	77.73%	77.40%	77.57%
overall	76.52%	74.82%	75.66%

Dutch test	precision	recall	$F_{\beta=1}$
LOC	78.54%	80.67%	79.59%
MISC	81.03%	70.51%	75.41%
ORG	70.03%	72.75%	71.36%
PER	80.71%	82.25%	81.47%
overall	77.83%	76.29%	77.05%