

Statistical Methods for Named Entity Recognition

Lluís Màrquez

TALP Research Center, Software Department
Universitat Politècnica de Catalunya

IULA, December 16, 2005

Document composed using: **pdflatex**, **ppower4**, **xfig** (with multi meta post format), **mpost**

Outline of the Talk

- Named Entity Recognition
- Standard Statistical Approach to NER
- Machine Learning Approach to NER
- Some Challenges
 - ★ Scarcity of Resources: Multilinguality
 - ★ Scarcity of Resources: Bootstrapping
 - ★ Joint Learning of NE and Relations
 - ★ Difficult domains: speech corpora

Outline of the Talk

- **Named Entity Recognition**
- Standard Statistical Approach to NER
- Machine Learning Approach to NER
- Some Challenges
 - ★ Scarcity of Resources: Multilinguality
 - ★ Scarcity of Resources: Bootstrapping
 - ★ Joint Learning of NE and Relations
 - ★ Difficult domains: speech corpora

Natural Language Processing Problems

Named Entity Recognition

Wolff, currently a journalist in Argentina, played with Del Bosque in the final years of the seventies in Real Madrid.

Natural Language Processing Problems

Named Entity Recognition

[**PER** Wolff] , currently a journalist in [**LOC** Argentina] ,
played with [**PER** Del Bosque] in the final years of the
seventies in [**ORG** Real Madrid] .

Natural Language Processing Problems

Named Entity Recognition

[**PER** Wolff] , currently a journalist in [**LOC** Argentina] , played with [**PER** Del Bosque] in the final years of the seventies in [**ORG** Real Madrid] .

- Structurally NER is similar to any other base “phrase”-recognition NLP task, e.g., NP-chunking
- But NER is a more semantic task (also with very strong orthographic and lexical cues)

Natural Language Processing Problems

Shallow Parsing: NP-Chunking

He reckons [NP the current account deficit] will narrow to [NP only 1.8 billion] in [NP September] .

Can be redefined as a sequential labeling problem

He_O reckons_O the_B-NP current_I-NP account_I-NP deficit_I-NP will_O narrow_O to_O only_B-NP 1.8_I-NP billion_I-NP in_O September_B-NP ._O

Natural Language Processing Problems

Named Entity Recognition

[**PER** Wolff] , currently a journalist in [**LOC** Argentina] , played with [**PER** Del Bosque] in the final years of the seventies in [**ORG** Real Madrid] .

- **Utility?**
- **Why learning?**
- **Difficulties?**

Natural Language Processing Problems

Named Entity Recognition

- Extensions
 - ★ Named Entities may be embedded
 - ★ NE tracing: variants and co-reference resolution
 - ★ Relations between entities: event extraction

Outline of the Talk

- Named Entity Recognition
- **Standard Statistical Approach to NER**
- Machine Learning Approach to NER
- Some Challenges
 - ★ Scarcity of Resources: Multilinguality
 - ★ Scarcity of Resources: Bootstrapping
 - ★ Joint Learning of NE and Relations
 - ★ Difficult domains: speech corpora

Statistical/Probabilistic Methods

- Evaluations at MUC conferences (mid 90s)
- Methods based on HMMs
- More sophisticated probabilistic methods: MEMM, CRF, etc. (2003–)

Statistical/Probabilistic Methods

- Evaluations at MUC conferences (mid 90s)
- Methods based on HMMs
- More sophisticated probabilistic methods: MEMM, CRF, etc. (2003–)
- **IdentiFinderTM**

Outline of the Talk

- Named Entity Recognition
- Standard Statistical Approach to NER
- **Machine Learning Approach to NER**
- Some Challenges
 - ★ Scarcity of Resources: Multilinguality
 - ★ Scarcity of Resources: Bootstrapping
 - ★ Joint Learning of NE and Relations
 - ★ Difficult domains: speech corpora

Machine Learning-based Named Entity Recognition

- Shared Tasks at the Conference on Natural Language Learning (CoNLL; 2002, 2003)
- Machine Learning for sequential tagging
- Many different learning algorithms for local decisions (DTs, AdaBoost, SVM, TiMBL, TBL, etc.)

Machine Learning-based Named Entity Recognition

- Shared Tasks at the Conference on Natural Language Learning (CoNLL; 2002, 2003)
- Machine Learning for sequential tagging
- Many different learning algorithms for local decisions (DTs, AdaBoost, SVM, TiMBL, TBL, etc.)
- **TALP system at CoNLL-2002 shared task**

Outline of the Talk

- Named Entity Recognition
- Standard Statistical Approach to NER
- Machine Learning Approach to NER
- **Some Challenges**
 - ★ **Scarcity of Resources: Multilinguality**
 - ★ Scarcity of Resources: Bootstrapping
 - ★ Joint Learning of NE and Relations
 - ★ Difficult domains: speech corpora

Outline of the Talk

- Named Entity Recognition
- Standard Statistical Approach to NER
- Machine Learning Approach to NER
- **Some Challenges**
 - ★ Scarcity of Resources: Multilinguality
 - ★ **Scarcity of Resources: Bootstrapping**
 - ★ Joint Learning of NE and Relations
 - ★ Difficult domains: speech corpora

Outline of the Talk

- Named Entity Recognition
- Standard Statistical Approach to NER
- Machine Learning Approach to NER
- **Some Challenges**
 - ★ Scarcity of Resources: Multilinguality
 - ★ Scarcity of Resources: Bootstrapping
 - ★ **Joint Learning of NE and Relations**
 - ★ Difficult domains: speech corpora

Outline of the Talk

- Named Entity Recognition
- Standard Statistical Approach to NER
- Machine Learning Approach to NER
- **Some Challenges**
 - ★ Scarcity of Resources: Multilinguality
 - ★ Scarcity of Resources: Bootstrapping
 - ★ Joint Learning of NE and Relations
 - ★ **Difficult domains: speech corpora**

Thank you very much for your attention!

Wide-Coverage Spanish Named Entity Extraction

Xavier Carreras, Lluís Màrquez & Lluís Padró



TALP Research Center

LSI, UPC
Technical University of Catalonia

IBERAMIA'02

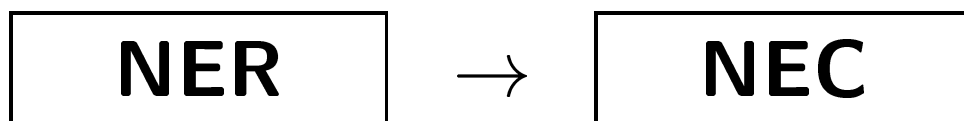
Wide-Coverage Spanish Named Entity Extraction

Named Entity Extraction

- The problem:

“Según informó el (**Departamento**)_{ORG} que dirige el consejero (**Inaxio Oliveri**)_{PER} , los representantes de la (**Consejería**)_{ORG} y de las universidades del (**País Vasco**)_{LOC}, (**Deusto**)_{ORG} y (**Mondragón**)_{ORG} estudiaron los nuevos retos de estos centros educativos.”

- Useful for: Information extraction, improving linguistic processors, automatic summarization, document linking/clustering, etc.
- Our Approach: Named Entity Extraction as two separate modules:



- **NER** = sequence tagging (e.g., IOB tagging)
- **NEC** = classification task

Named Entity Extraction

- All (local) decisions resolved with Machine Learning–based classifiers
- Learning Algorithm: **AdaBoost**
 - Combination of many *weak* classifiers (hypotheses or rules) into a *strong* classifier
 - Binary classification, binary features
 - Real-valued AdaBoost with confidence-rated predictions (Schapire & Singer 99)
 - Combined classifier: $f(x) = \sum_{t=1}^T \alpha_t h_t(x)$
 - Weak Rules (h_t): Decision Trees of fixed depth
 - Good performance/efficiency in NLP domains (Abney et al., 99; Schapire & Singer, 00; Escudero et al., 00) (Carreras & Màrquez, 01; Carreras et al., 02a; 02b; etc.)

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

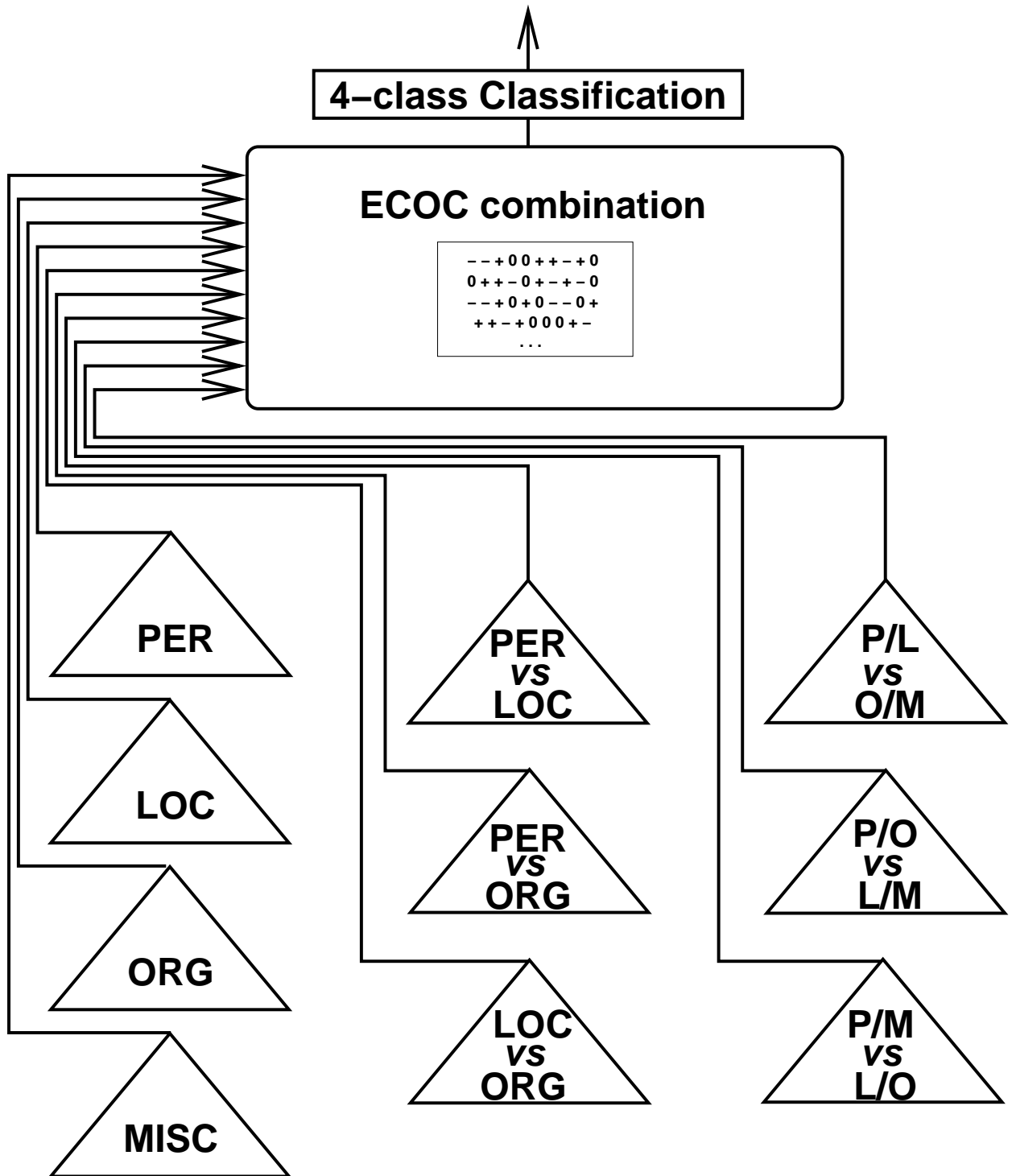
- greedy Open-Close, plus Inside checking:

	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

NEC Module



Features

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

<i>initial-caps</i>	<i>all-caps</i>	<i>all-digits</i>
<i>roman-number</i>	<i>contains-dots</i>	<i>contains-hyphen</i>
<i>acronym</i>	<i>lonely-initial</i>	<i>punctuation-mark</i>
<i>single-char</i>	<i>functional-word</i>	<i>URL</i>

- 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 (BIO or Open/Close tags)

Evaluation

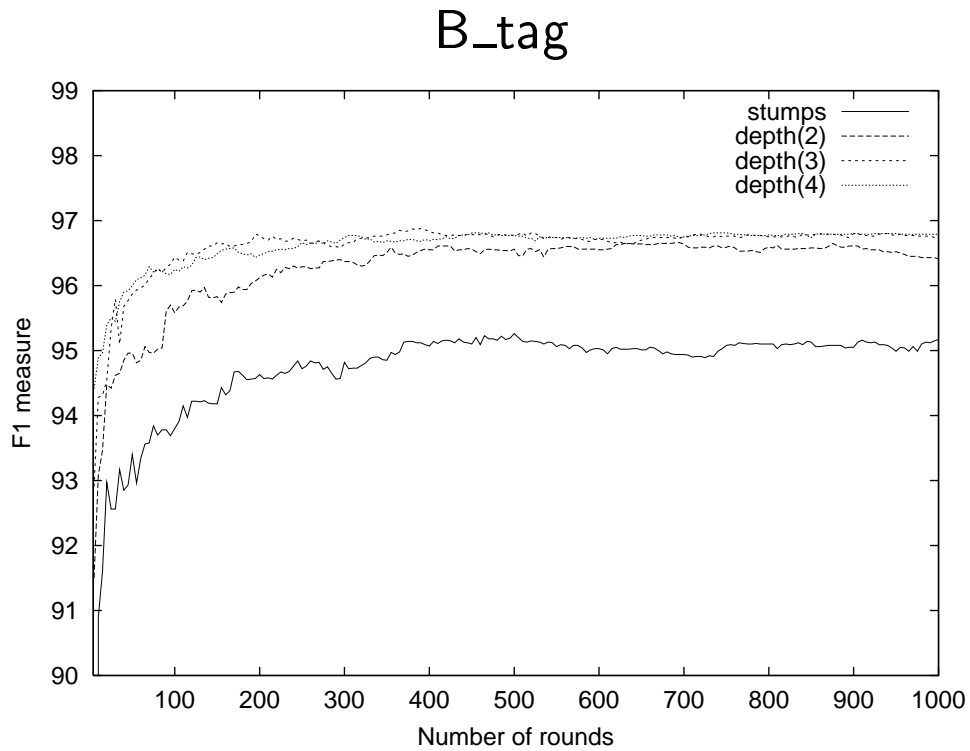
- The(*Agencia*) EFE Spanish corpus:
 - Collection of over 3,000 news agency articles issued during year 2000
 - 802,729 words, which contain about 86,000 hand tagged named entities
- Experimental setting:
 - Test subset: $\sim 65,000$ words (4,820 NE)
 - NEC training set: $\sim 700,000$ words (61,266 NE)
 - NER training set: $\sim 100,000$ words (8,126 NE)
 - Only features occurring more than 2 times
 - Learning parameters: *number of rounds, depth of the base classifiers* (from stumps to decision trees of depth 4)
 - Evaluation measures: *precision, recall, F_1 , accuracy*

Learning Problems: Sizes

NER	#Exs.	#Feat.	#Pos.examples
open	91,625	19,215	8,126 (8.87%)
close	8,802	10,795	4,820 (54.76%)
I	97,333	20,526	5,708 (5.86%)
O	97,333	20,526	83,499 (85.79%)
B	97,333	20,526	8,126 (8.35%)

NEC	#Exs.	#Feat.	#Pos.examples
PERSON	61,266	22,465	12,976 (21.18%)
LOCATION	61,266	22,465	14,729 (24.04%)
ORGANIZ	61,266	22,465	22,947 (34.46%)
OTHER	61,266	22,465	10,614 (17.32%)

Results on the NER Task (1)



Method	P	R	F_1
MACO+	89.94%	87.51%	88.71%
OpenClose	92.42%	91.54%	91.97%
BIO	92.66%	91.99%	92.33%
OpenClose+I	92.60%	92.14%	92.37%

Results on the NER Task (2)

Subset	#NE	<i>P</i>	<i>R</i>	<i>F</i>₁
length=1	2,807	94.64%	95.65%	95.15%
length=2	1,005	94.01%	93.73%	93.87%
length=3	495	91.65%	88.69%	90.14%
length=4	237	84.81%	84.81%	84.81%
length=5	89	77.27%	76.40%	76.84%
length=6	74	81.94%	79.73%	80.82%
length=7	22	60.00%	54.55%	57.14%
length=8	22	88.24%	68.18%	76.92%
length=9	11	80.00%	72.73%	76.19%
length=10	3	50.00%	33.33%	40.00%
uppercase	4,637	92.81%	93.25%	93.03%
lowercase	183	85.40%	63.93%	73.13%
TOTAL	4,820	92.60%	92.14%	92.37%

Results on the NEC Task

Method	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Acc</i>
Most freq	39.78%	39.78%	39.78%	39.78%
basic	90.19%	84.44%	87.22%	87.51%
+tw	90.11%	84.77%	87.36%	88.17%
+gaz	90.25%	85.31%	87.71%	88.60%
+tw+gaz	90.61%	85.23%	87.84%	88.73%

Method	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Acc</i>
Most freq	37.47%	37.47%	37.47%	37.47%
basic	83.84%	79.39%	81.55%	81.85%
+tw	85.08%	79.30%	82.09%	82.15%
+gaz	85.05%	79.57%	82.22%	82.15%
+tw+gaz	85.32%	79.80%	82.47%	82.31%

Comparison to other Systems

- CoNLL'02 Shared Task on NERC
 - Organized by the ACL's SIG on NLL (Sep.02)
 - Spanish and Dutch datasets publicly available
 - 12 participants using ML-based systems
 - Best results in both languages (~ 2 points F_1)
 - Why?
 - * Appropriateness of the learning algorithm
 - * Design of the learning attributes
 - * Combination of classifiers

Conclusions and Further Work

- Conclusions:
 - NERC system for Spanish based on robust Machine Learning techniques
 - Large set of simple features requiring no complex linguistic processing
 - Fairly good performance (validated in the CoNLL'02 competition framework)
 - Gazetteers and trigger words allow to slightly improve performance
- Current and future work:
 - Classification algorithms other than AdaBoost (e.g., SVMs)
 - Use of global inference schemes
 - Multi-label AdaBoost algorithms
 - Reusing of the Spanish models to develop a system for Catalan (work submitted to EACL)

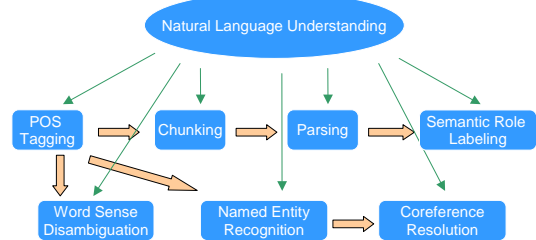
A Linear Programming Formulation for Global Inference in Natural Language Tasks

Dan Roth Wen-tau Yih (yih@uiuc.edu)

Department of Computer Science
University of Illinois at Urbana-Champaign

Page 1

View of Solving NLP Problems



Page 2

Weaknesses of Pipeline Model

- Propagation of errors
- Bi-Directional interactions between stages
 - Occasionally, later stage problems are easier.

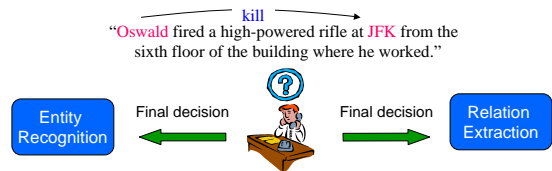
“Oswald fired a high-powered rifle at JFK from the sixth floor of the building where he worked.”

- Upstream mistakes will not be corrected.

Page 3

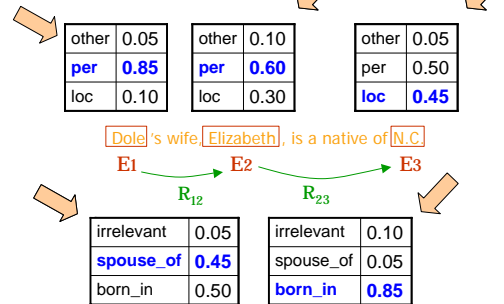
Global Inference with Classifiers

- Classifiers (for components) are trained or given in advance.
- There are constraints on classifiers' labels (which may be known during training or only known during testing).
- The inference procedure attempts to make the best global assignment, given the local predictions and the constraints



Page 4

Ideal Inference



Page 5

Inference Procedure

- Inference with classifiers is not a new idea.
 - On sequential constraint structure:
 - HMM, PMM, CRF[Lafferty et al.], CSCL[Punyakanok&Roth]
 - On general structure: Heuristic search
- Integer linear programming (ILP) formulation
 - General: works on non-sequential constraint structure
 - Flexible: can represent many types of constraints
 - Optimal: finds the optimal solution
 - Fast: commercial packages are able to solve it quickly

Page 6

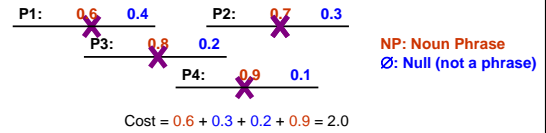
Outline

- Case study – using ILP to solve:
 1. Non-overlapping constraints
 - Usually solved using dynamic programming on sequential constraint structure
 2. Simultaneous entity/relation recognition
 - ILP can go beyond sequential constraint structure
- Discussion – pros & cons
- Summary & Future Work

Page 7

Phrase Identification Problems

- Several NLP problems attempt to detect and classify phrases in sentences.
 - e.g., Named Entity Recognition, Shallow Parsing, Information Extraction, Semantic Role Labeling, etc.
- Given a sentence, classifiers (OC, or phrase) predict phrase candidates (which often overlap).



Page 8

LP Formulation – Linear Cost

- Indicator variables

$$x_{\{P1=NP\}}, x_{\{P1=\emptyset\}}, \dots, x_{\{P4=NP\}}, x_{\{P4=\emptyset\}} \in \{0,1\}$$

Total Cost

$$= c_{\{P1=NP\}} \cdot x_{\{P1=NP\}} + c_{\{P1=\emptyset\}} \cdot x_{\{P1=\emptyset\}} + \dots + c_{\{P4=\emptyset\}} \cdot x_{\{P4=\emptyset\}}$$

$$= 0.6 \cdot x_{\{P1=NP\}} + 0.4 \cdot x_{\{P1=\emptyset\}} + \dots + 0.1 \cdot x_{\{P4=\emptyset\}}$$

Page 9

LP Formulation – Linear Constraints

Subject to:

Non-overlapping Constraints:

$$x_{\{P1=\emptyset\}} + x_{\{P3=\emptyset\}} \geq 1$$

$$x_{\{P2=\emptyset\}} + x_{\{P3=\emptyset\}} + x_{\{P4=\emptyset\}} \geq 2$$

Page 10

LP Formulation

$$\max \sum_{i \in Phrases, k \in Classes} c_{i,k} \cdot x_{i,k}$$

subject to:

$$x_{i,k} \in \{0,1\} \quad \forall i \in Phrases, k \in Classes$$

$$\sum_{k \in Classes} x_{i,k} = 1 \quad \forall i \in Phrases$$

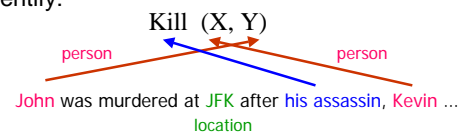
$$\sum_{i=1}^m x_{j_i,k} = m - 1 \quad \forall P_{j_i}, 1 \leq i \leq m, \text{ that overlap}$$

Generate one integer linear program per sentence.

Page 11

Entity/Relation Recognition

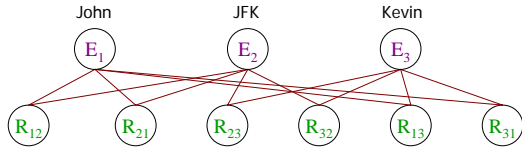
John was murdered at JFK after his assassin, Kevin...
Identify:



- Identify named entities
- Identify relations between entities
- Exploit mutual dependencies between named entities and relations to yield a coherent global prediction

Page 12

Problem Setting

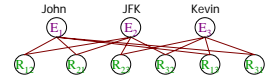


Constraints:

- $(R_{12} = \text{kill}) \rightarrow (E_1 = \text{person}) \wedge (E_2 = \text{person})$
- $(R_{12} = \text{headquarter}) \rightarrow (E_1 = \text{organization}) \wedge (E_2 = \text{location})$
- ...

Page 13

LP Formulation – Indicator Variables



- For each variable

$$x_{\{E1 = \text{per}\}}, x_{\{E1 = \text{loc}\}}, \dots, \\ x_{\{R12 = \text{kill}\}}, x_{\{R12 = \text{born_in}\}}, \dots, x_{\{R12 = \emptyset\}}, \dots \in \{0,1\}$$

- For each pair of variables on an edge

$$x_{\{R12 = \text{kill}, E1 = \text{per}\}}, x_{\{R12 = \text{kill}, E1 = \text{loc}\}}, \dots, \\ x_{\{R12 = \emptyset, E1 = \text{per}\}}, x_{\{R12 = \emptyset, E1 = \text{loc}\}}, \dots, \\ x_{\{R32 = \emptyset, E2 = \text{per}\}}, x_{\{R32 = \emptyset, E2 = \text{loc}\}}, \dots \in \{0,1\}$$

Page 14

LP Formulation – Cost Function

- Assignment cost

$$c_{\{E1 = \text{per}\}} \cdot x_{\{E1 = \text{per}\}} + c_{\{E1 = \text{loc}\}} \cdot x_{\{E1 = \text{loc}\}} + \dots + \\ c_{\{R12 = \text{kill}\}} \cdot x_{\{R12 = \text{kill}\}} + \dots + c_{\{R12 = \emptyset\}} \cdot x_{\{R12 = \emptyset\}} + \dots$$

- Constraint cost

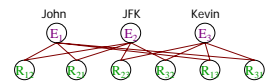
$$c_{\{R12 = \text{kill}, E1 = \text{per}\}} \cdot x_{\{R12 = \text{kill}, E1 = \text{per}\}} + \\ -\infty \rightarrow c_{\{R12 = \text{kill}, E1 = \text{loc}\}} \cdot x_{\{R12 = \text{kill}, E1 = \text{loc}\}} + \dots + \\ c_{\{R12 = \emptyset, E1 = \text{loc}\}} \cdot x_{\{R12 = \emptyset, E1 = \text{loc}\}} + \dots$$

- Total cost = Assignment cost + Constraint cost

Page 15

LP Formulation – Linear Constraints

Subject to:



Node–Edge Consistency Constraints

$$x_{\{R12 = \text{kill}\}} = x_{\{R12 = \text{kill}, E1 = \text{Null}\}} + \\ x_{\{R12 = \text{kill}, E1 = \text{person}\}} + \\ x_{\{R12 = \text{kill}, E1 = \text{location}\}} + \\ x_{\{R12 = \text{kill}, E1 = \text{organization}\}}$$

Page 16

LP Formulation

$$\max \sum_{i \in \mathcal{E}, k \in \mathcal{L}_{\mathcal{E}}} c_{i,k} \cdot x_{i,k} + \sum_{i \in \mathcal{R}, k \in \mathcal{L}_{\mathcal{R}}} c_{i,k} \cdot x_{i,k} + \sum_{r \in \mathcal{R}, e \in \mathcal{E}, h \in \mathcal{L}_{\mathcal{R}}, r \in \mathcal{L}_{\mathcal{E}}} c_{r,h,e} \cdot x_{r,h,e}$$

subject to:

$$x_{E,e} \in \{0,1\} \quad \forall E \in \mathcal{E}, e \in \mathcal{L}_{\mathcal{E}} \\ x_{R,r} \in \{0,1\} \quad \forall R \in \mathcal{R}, r \in \mathcal{L}_{\mathcal{R}} \\ x_{H,r,E,e} \in \{0,1\} \quad \forall R \in \mathcal{R}, r \in \mathcal{L}_{\mathcal{R}}, E \in \mathcal{E}, e \in \mathcal{L}_{\mathcal{E}} \\ \sum_{e \in \mathcal{L}_{\mathcal{E}}} x_{\{E,e\}} = 1 \quad \forall E \in \mathcal{E} \\ \sum_{r \in \mathcal{L}_{\mathcal{R}}} x_{\{H,r\}} = 1 \quad \forall R \in \mathcal{R} \\ x_{E,e} = \sum_{r \in \mathcal{L}_{\mathcal{R}}} x_{R,r,E,e} \quad \forall E \in \mathcal{E} \text{ and } \forall R \in \mathcal{N}(E) \\ x_{H,r} = \sum_{e \in \mathcal{L}_{\mathcal{E}}} x_{H,r,E,e} \quad \forall R \in \mathcal{R} \text{ and } \forall E \in \mathcal{N}(R)$$

Generate one integer linear program per sentence.

Page 17

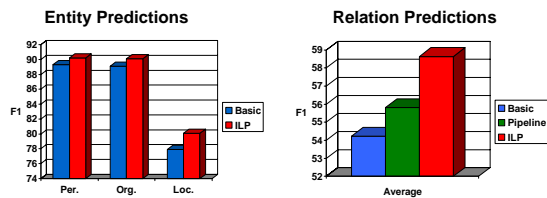
Experiments – Data

Methodology: 1,437 sentences from TREC data;
5,336 entities; 19,048 pairs of potential relations.

Relation	Entity1	Entity2	Example	#
Located-in	Loc	Loc	(New York, US)	406
Work-for	Per	Org	(Bill Gates, Microsoft)	394
OrgBased-in	Org	Loc	(HP, Palo Alto)	451
Live-in	Per	Loc	(Bush, US)	521
Kill	Per	Per	(Oswald, JFK)	268

Page 18

Experimental Results – F_1

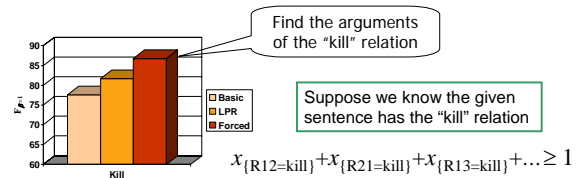


- Improvement compared to the **basic** (w/o inference) and **pipeline** (entity→relation) models
- Quality of decisions is enhanced
 - No “stupid mistakes” that violate global constraints

Page 19

Decision-time Constraint

- Constraints may be known only in decision time.
 - Question Answering: “Who killed JFK?”
 - Find “kill” relation in candidate sentences



Page 20

Computational Issues

- Exhaustive search won't work
 - Even with a small number of variables and classes, the solution space is intractable
 - $n=20, k=5, 5^{20} = 95,367,431,640,625$
- Heuristic search algorithms (e.g., beam search)?
 - Do not guarantee optimal solutions
 - In practice, may not be faster than ILP

Page 21

Generality (1/2)

- Linear constraints can represent any Boolean function
 - More components can be put in this framework
 - Who killed whom? (determine arguments of the Kill relation)
 - Entity1=Entity3 (co-ref classifier)
 - Subj-Verb-Object constraints
- Able to handle non-sequential constraint structure
 - E/R case has demonstrated this property

Page 22

Generality (2/2)

- Integer linear programming (ILP) is NP-hard.
- However, an ILP problem at this scale can be solved very quickly using commercial packages, such as CPLEX or Xpress-MP.
 - CPLEX is able to solve a linear programming problem of 13 million variables within 5 minutes.
 - Processing 20 sentences in a second for a named entity recognition task on P3-800MHz

Page 23

Current/Future Work

- Handle stochastic (soft) constraints
 - Example: If the relation is *kill*, the first argument is *person* with 0.95 probability, and *organization* with 0.05 probability.
- Incorporate inference at learning time
 - Along the lines of [Carreras & Marquez, NIPS-03]

Page 24

Named Entity Recognition from Spontaneous Open-Domain Speech

Mihai Surdeanu, Jordi Turmo, and Eli Comelles

TALP Research Center

Universitat Politècnica de Catalunya, Barcelona, Spain

{surdeanu,turmo,comelles}@lsi.upc.edu

INTERSPEECH - 2005

Goal of the work

- This paper presents an analysis of named entity recognition and classification in spontaneous speech transcripts.
- A significant fraction of the Switchboard corpus is annotated with six named entity classes and investigated a battery of machine learning models that include lexical, syntactic, and semantic features
- The best recognition and classification model obtains promising results, approaching within 5% a system evaluated on clean textual data

The Switchboard Corpus

The corpus used in this paper is an extension of the Switchboard (SWB) corpus (LDC catalog number LDC97S62).

B: You mean they don't have the, uh, the smog alerts?
A: No, not in, not in Te-, well not in Dallas, that is.
B: Right. I, I,
A: [throat_clearing].
B: yeah, I spent a summer i-, i-, in Tyler so I know, just east of Dallas there
A: Yeah. We're going there tomorrow.

Sample SWB transcript fragment with two speakers.

The Switchboard Corpus

- Spontaneous speech difficulties
 - disfluency or stuttering
 - speaker corrections and specifications
 - lack of grammatical structure
 - lack of case information

Features Models under Study

- **Model M1** – contains only lexical attributes: *lemma*, *prefixes*, *suffixes*, *case information*, etc.
- **Model M2** – adds format attributes: *all-caps*, *caps+dots*, *all-digits*, etc.
- **Model M3** – adds part of speech (POS) attributes (using TnT)
- **Model M4** – adds basic syntactic phrases (chunks) attributes (using YamCha)

Features Models under Study

- **Model M5** – adds more elaborate syntactic context features: headwords of neighboring chunks.
- **Model M6** – adds class-based attributes: *is-number*, *is-day*, *is-month*, *is-multiplier*, etc.
- **Model M7** – adds gazetteer-based attributes: first and last person names and locations.

Machine Learning Applied

- NER+NEC decomposition
- NER: BIO sequential tagging with left-recurrences
- NER and NEC trained with SVMs with degree 2 polynomial kernels

Experimental Results

(1) F measure on SWB transcripts *with* case information

	M1	M2	M3	M4	M5	M6	M7
LOC	81.13	83.50	83.92	83.34	82.71	82.18	83.11
MISC	56.14	65.57	65.39	65.24	65.48	66.49	66.36
MONEY	76.60	74.61	75.00	75.79	81.22	80.39	82.59
ORG	62.63	63.33	64.21	64.97	64.25	64.16	64.83
PER	62.48	74.27	72.73	70.76	69.22	70.30	77.72
TIME	76.68	75.50	75.90	76.00	76.24	75.88	75.37
Overall	70.78	74.24	74.32	74.04	73.83	73.96	75.12

Best model = M1 + M2 + M3 + M6 + M7

Experimental Results

(2) F measure on SWB transcripts *without* case information

	M1	M2	M3	M4	M5	M6	M7
LOC	80.86	81.29	80.15	80.27	79.07	79.06	79.90
MISC	53.89	54.62	53.42	53.15	48.81	49.37	50.88
MONEY	78.31	78.72	77.49	76.84	79.59	79.00	79.41
ORG	61.79	62.68	62.68	63.19	58.71	59.85	58.10
PER	62.72	65.07	67.88	63.22	59.48	61.51	69.80
TIME	76.13	75.53	75.50	75.80	74.75	75.89	74.45
Overall	70.09	70.56	70.09	69.67	67.55	68.13	69.22

Best model = M1 + M2 + M6 + M7

Experimental Results

(3) F measure of the best models on written text versus speech

Written Text (CoNLL)	Speech (SWB with case)	Speech(SWB without case)
80.52	75.50	71.55

(4) Justification for the Spontaneous-Speech Corpus: F measure drop when training the best NERC model on "clean" textual data and testing on SWB.

LOC	MISC	ORG	PER	Overall
-13.12	-43.45	-30.66	-22.19	-24.68

Named Entity Recognition from Spontaneous Open-Domain Speech

Mihai Surdeanu, Jordi Turmo, and Eli Comelles

TALP Research Center

Universitat Politècnica de Catalunya. Barcelona, Spain

{surdeanu,turmo,comelles}@lsi.upc.edu

INTERSPEECH - 2005

Named Entity Recognition for Catalan Using Spanish Resources

Xavier Carreras, Lluís Màrquez and Lluís Padró

TALP Research Center

LSI Department

Universitat Politècnica de Catalunya

{carreras,lluism,padro}@lsi.upc.es



EACL-2003, Budapest, April 15-17

Outline

- Introduction
- Resources/Tools
- Approaches
- Evaluation
- Bootstrapping
- Conclusions



EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Outline

- **Introduction**
- Resources/Tools
- Approaches
- Evaluation
- Bootstrapping
- Conclusions



EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Introduction

Introduction

ML-based Named Entity Recognition for Catalan

- Building a low-cost ML-based system
- No available resources
- Use existing Spanish resources
- Take advantage of their similarities:
 - NE internal structure
 - NE context structure
 - Social and cultural environments



EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Introduction

An Example

Spanish

"El presidente del **Comité Olímpico Internacional**, **José Antonio Samaranch**, se reunió el lunes en Nueva York con investigadores del **FBI** y del **Departamento de Justicia**."

Catalan

"El president del **Comitè Olímpic Internacional**, **Josep Antoni Samaranch**, es va reunir dilluns a Nova York amb investigadors del **FBI** i del **Departament de Justícia**."



EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Introduction

An Example

Spanish

"El presidente del **Comité Olímpico Internacional**, **José Antonio Samaranch**, se reunió el lunes en Nueva York con investigadores del **FBI** y del **Departamento de Justicia**."

Catalan

"El president del **Comitè Olímpic Internacional**, **Josep Antoni Samaranch**, es va reunir dilluns a Nova York amb investigadors del **FBI** i del **Departament de Justícia**."



EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Introduction

An Example

Spanish

"El presidente del **Comité Olímpico Internacional**, **José Antonio Samaranch**, se reunió el lunes en **Nueva York** con investigadores del **FBI** y del **Departamento de Justicia**."

Catalan

"El president del **Comitè Olímpic Internacional**, **Josep Antoni Samaranch**, es va reunir dilluns a **Nova York** amb investigadors del **FBI** i del **Departament de Justícia**."

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Introduction

An Example

Spanish

"El presidente del **Comité Olímpico Internacional**, **José Antonio Samaranch**, se reunió el lunes en **Nueva York** con investigadores del **FBI** y del **Departamento de Justicia**."

Catalan

"El president del **Comitè Olímpic Internacional**, **Josep Antoni Samaranch**, es va reunir dilluns a **Nova York** amb investigadors del **FBI** i del **Departament de Justícia**."

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Outline

- Introduction
- **Resources/Tools**
- Approaches
- Evaluation
- Bootstrapping
- Conclusions

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Resources

Existing Spanish Resources

- CoNLL-2002 corpus:
 - From the EFE Newswire Agency collection
 - ~365,000 words. ~26,000 NEs
 - Hand-tagged
 - Divided into **train** (~70%), **development** and **test** sets (~15% each)
- AdaBoost-based recognition models for the CoNLL-2002 Spanish NER task (Carreras et al., 02)

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Resources

Spanish NER Module

(Carreras et al., 2002)

- Sequential B-I-O tagging
- AdaBoost binary classifiers
- Features: Window-based
 - **Lexical**: *word forms*.
 - **Orthographic**: *capitalization, digits, hyphenation, etc.*
 - **Affixes**: *prefixes, suffixes (up to 4 letters)*
 - **Word-type patterns**: *functional, capitalized, lowercase, punctuation, quote, other.*
 - **Left predictions**

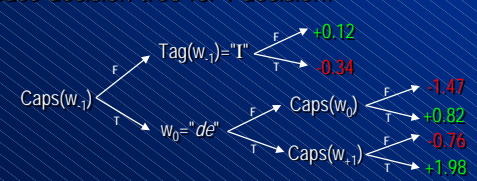
EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Resources

Spanish NER Module

(Carreras et al., 2002)

- AdaBoost constructs an ensemble of base/weak classifiers, which are linearly combined for making real-valued predictions (Schapire & Singer, 1999)
- Base decision tree for I decision:

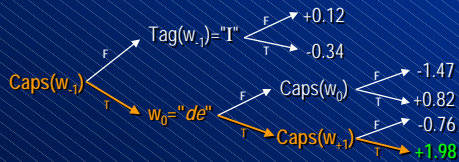


EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Spanish NER Module

(Carreras et al., 2002)

- AdaBoost constructs an ensemble of base/weak classifiers, which are linearly combined for making real-valued predictions (Schapire & Singer, 1999)
- Base decision tree for 1 decision:



EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Existing Catalan Resources

- Catalan corpus:
 - From the daily newspaper: *"El Periódico de Catalunya"*
 - Subset of ~2,200,000 words from the beginning of year 2,000
 - No linguistic annotation

EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Outline

- Introduction
- Spanish Resources/Tools
- **Approaches**
- Evaluation
- Bootstrapping
- Conclusions

EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Approaches

- **Classical**: Hand tagging a Catalan training corpus
- **Straight**: Use existing Spanish AdaBoost models straightforwardly on Catalan data
- **Translation**: Translate existing Spanish AdaBoost models
- **X-Ling**: Use existing Spanish (+Catalan) resources to train a Spanish-Catalan bilingual model. Cross-linguistic features.

EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Approaches: Classical

Hand tagging a Catalan training corpus

- Training set: 23,177 words containing 1,232 NEs
- Test set: 23,595 words containing 1,338 NEs
- Annotation labour cost: **10 person hours**

EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Approaches: Straight

Use existing Spanish Adaboost models on Catalan data

- Models learnt on existing Spanish training data: 264,715 words containing 18,797 NEs
- Lexical features removed to reduce noise
- Labour cost: **0 person hours**

EACL-2003, Budapest, 15-17 April 2003

16/4/2003

Approaches: Translation

Translate existing Spanish AdaBoost models

- Models acquired on existing Spanish training data: 264,715 words containing 18,797 NEs
- Translation of lexical features (<5,000), e.g.:
 $[w_j = \text{"calle"}] \Rightarrow [w_j = \text{"carrer"}]$
- Translation labour cost:
 - 10 person hours ("common-sense" hand translation)
 - 0.5 person hours (InterNOSTRUM automatic translation + shallow supervision)



Approaches: X-Ling

Use existing Spanish corpora to train a Catalan model

- Compilation of a dictionary of translation pairs
- Makes use of bilingual features, e.g.:
 $[(w_j = \text{calle} \wedge \text{lang} = \text{es}) \vee (w_j = \text{carrer} \wedge \text{lang} = \text{ca})]$
- Systems may be trained on any of both languages, either separately or jointly
- They may be used on any of both languages



Outline

- Introduction
- Spanish Resources/Tools
- Approaches
- **Evaluation**
- Bootstrapping
- Conclusions



Evaluation

	train	Feature transl.	es test			ca test		
			prec.	rec.	F ₁	prec.	rec.	F ₁
Classical	ca	—	—	—	—	90.98	87.44	89.18
Straight	es	—	89.31	88.03	88.67	82.80	82.21	82.50
Translation	es	man.	92.81	92.89	92.85	89.14	92.00	90.55
		aut.	83.85	91.55	87.53	83.85	91.55	87.53
X-Ling (es)	es	man.	92.25	92.64	92.44	90.78	89.76	90.27
		aut.	92.23	92.69	92.46	89.95	89.61	89.78
X-Ling (mix)	es+ca	man.	92.27	92.53	92.40	91.95	90.43	91.18
		aut.	92.57	92.39	92.48	91.29	90.13	90.71



Evaluation

	train	Feature transl.	es test			ca test		
			prec.	rec.	F ₁	prec.	rec.	F ₁
Classical	ca	—	—	—	—	90.98	87.44	89.18
Straight	es	—	89.31	88.03	88.67	82.80	82.21	82.50
Translation	es	man.	92.81	92.89	92.85	89.14	92.00	90.55
		aut.	83.85	91.55	87.53	83.85	91.55	87.53
X-Ling (es)	es	man.	92.25	92.64	92.44	90.78	89.76	90.27
		aut.	92.23	92.69	92.46	89.95	89.61	89.78
X-Ling (mix)	es+ca	man.	92.27	92.53	92.40	91.95	90.43	91.18
		aut.	92.57	92.39	92.48	91.29	90.13	90.71



Evaluation

	train	Feature transl.	es test			ca test		
			prec.	rec.	F ₁	prec.	rec.	F ₁
Classical	ca	—	—	—	—	90.98	87.44	89.18
Straight	es	—	89.31	88.03	88.67	82.80	82.21	82.50
Translation	es	man.	92.81	92.89	92.85	89.14	92.00	90.55
		aut.	83.85	91.55	87.53	83.85	91.55	87.53
X-Ling (es)	es	man.	92.25	92.64	92.44	90.78	89.76	90.27
		aut.	92.23	92.69	92.46	89.95	89.61	89.78
X-Ling (mix)	es+ca	man.	92.27	92.53	92.40	91.95	90.43	91.18
		aut.	92.57	92.39	92.48	91.29	90.13	90.71



Evaluation

Evaluation

	train	Feature transl.	es test			ca test		
			prec.	rec.	F_1	prec.	rec.	F_1
Classical	ca	—	—	—	—	90.98	87.44	89.18
Straight	es	—	89.31	88.03	88.67	82.80	82.21	82.50
Translation	es	man.	92.81	92.89	92.85	89.14	92.00	90.55
		aut.	83.85	91.55	87.53	83.85	91.55	87.53
X-Ling (es)	es	man.	92.25	92.64	92.44	90.78	89.76	90.27
		aut.	92.23	92.69	92.46	89.95	89.61	89.78
X-Ling (mix)	es+ca	man.	92.27	92.53	92.40	91.95	90.43	91.18
		aut.	92.57	92.39	92.48	91.29	90.13	90.71

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Evaluation

Evaluation

	train	Feature transl.	es test			ca test		
			prec.	rec.	F_1	prec.	rec.	F_1
Classical	ca	—	—	—	—	90.98	87.44	89.18
Straight	es	—	89.31	88.03	88.67	82.80	82.21	82.50
Translation	es	man.	92.81	92.89	92.85	89.14	92.00	90.55
		aut.	83.85	91.55	87.53	83.85	91.55	87.53
X-Ling (es)	es	man.	92.25	92.64	92.44	90.78	89.76	90.27
		aut.	92.23	92.69	92.46	89.95	89.61	89.78
X-Ling (mix)	es+ca	man.	92.27	92.53	92.40	91.95	90.43	91.18
		aut.	92.57	92.39	92.48	91.29	90.13	90.71

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Evaluation

Evaluation

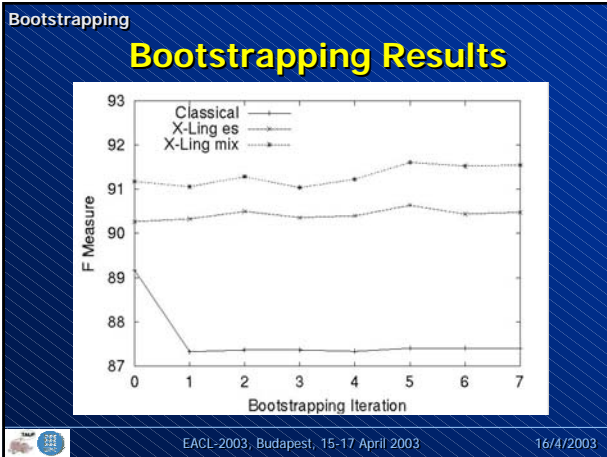
	train	Feature transl.	es test			ca test		
			prec.	rec.	F_1	prec.	rec.	F_1
Classical	ca	—	—	—	—	90.98	87.44	89.18
Straight	es	—	89.31	88.03	88.67	82.80	82.21	82.50
Translation	es	man.	92.81	92.89	92.85	89.14	92.00	90.55
		aut.	83.85	91.55	87.53	83.85	91.55	87.53
X-Ling (es)	es	man.	92.25	92.64	92.44	90.78	89.76	90.27
		aut.	92.23	92.69	92.46	89.95	89.61	89.78
X-Ling (mix)	es+ca	man.	92.27	92.53	92.40	91.95	90.43	91.18
		aut.	92.57	92.39	92.48	91.29	90.13	90.71

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

- Outline
- Introduction
 - Spanish Resources/Tools
 - Approaches
 - Evaluation
 - **Bootstrapping**
 - Conclusions
- EACL-2003, Budapest, 15-17 April 2003 16/4/2003

- Bootstrapping
- ## Bootstrapping
- Obtained models may be used in a bootstrapping process provided unannotated Catalan data is available.
 - Unannotated corpus: 2.2 Mwords, divided into $S_1 \dots S_N$ disjoint subsets of 1,000 sentences each.
 - Iteratively annotate a growing number of S subsets and retrain using all annotated corpus.
- EACL-2003, Budapest, 15-17 April 2003 16/4/2003

- Bootstrapping
- ## Bootstrapping Procedure
- M_0 =initial model
 T_L =initial labelled (training) corpus
for $i = 1 \dots N$ **do**
 - Identify NEs in $S_1 \dots S_i$ using model M_{i-1}
 - Learn a new model M_i using $T_L \cup \bigcup_{j=1}^i S_j$ as training data
endfor
output model M_N
- EACL-2003, Budapest, 15-17 April 2003 16/4/2003



Outline

- Introduction
- Spanish Resources/Tools
- Approaches
- Evaluation
- Bootstrapping
- **Conclusions**

EACL-2003, Budapest, 15-17 April 2003 16/4/2003

- Conclusions
- ## Conclusions
- A competitive low-cost Catalan NER system can be developed using existing Spanish resources
 - At the same cost, the hand translation of a Spanish model is better than the *classical* approach of learning from a small Catalan annotated corpus
 - The translation can be automatically done obtaining also very competitive results
- EACL-2003, Budapest, 15-17 April 2003 16/4/2003

- Conclusions
- ## Conclusions
- The best strategy turned out to be the use of cross linguistic features (*X-Ling* approach), which allows the training of models using mixed corpora and achieving good results on both languages
 - The *classical* model is not improved via bootstrapping, probably due to the small size of the Catalan training corpus
 - *X-Ling* models are slightly improved in the initial rounds of bootstrapping, producing robust models that do not degrade in further iterations
- EACL-2003, Budapest, 15-17 April 2003 16/4/2003

- Conclusions
- ## Current and Future work
- Extension to other similar Romance languages: Catalan, French, Galician, Italian, Spanish, etc. to train multilingual Named Entity recognition systems
 - Use a co-training based algorithm in order to improve the bootstrapping procedure: independent views (local-contextual features), independent NE taggers, different languages, etc.
 - To make a complete NERC system for Catalan
- EACL-2003, Budapest, 15-17 April 2003 16/4/2003

Named Entity Recognition for Catalan Using Spanish Resources

Xavier Carreras, Lluís Màrquez and Lluís Padró

TALP Research Center
LSI Department
Universitat Politècnica de Catalunya
{carreras,lluism,padro}@lsi.upc.es

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
UNIVERSITAT POLITÈCNICA DE CATALUNYA

EACL-2003, Budapest, April 15-17

An Algorithm that Learns What's in a Name (Identifinder™)

Dan M. Bikel, Richard Schwartz, Ralph Weischedel

Machine Learning

Special Issue on Natural Language Learning

C. Cardie and R. Mooney eds. 1999

Generalities

- HMM-based model for NERC
- Evolution from the *Nymble* system (1997)
- Applied in MUC conferences and other corpora with very good results
- Competitive (and better in some cases) to the best hand-developed rule-based NERC systems
- State-of-the-art for the task until CoNLL conferences (ML)

HMMs

sequence of observations : $\{o_1, \dots, o_n\}$

sequence of states : $\{s_1, \dots, s_n\}$

$$\arg \max_{s_1 \dots s_n} P(s_1, \dots, s_n | o_1, \dots, o_n) \approx$$

$$\arg \max_{s_1 \dots s_n} \prod_{k=1}^n P(s_k | s_{k-2}, s_{k-1}) \cdot P(o_k | s_k)$$

emission probabilities : $P(o_k | s_k)$

transition probabilities : $P(s_k | s_{k-2}, s_{k-1})$

initial state probabilities : $P(s_1)$

HMMs

sequence of observations : $\{o_1, \dots, o_n\}$

sequence of states : $\{s_1, \dots, s_n\}$

$$\arg \max_{s_1 \dots s_n} P(s_1, \dots, s_n | o_1, \dots, o_n) \approx$$

$$\arg \max_{s_1 \dots s_n} \prod_{k=1}^n P(s_k | s_{k-2}, s_{k-1}) \cdot P(o_k | s_k)$$

Decoding: The argmax function can be computed in linear time $O(n)$ using dynamic programming (Viterbi algorithm)

Identifinder HMM: Graphical Representation

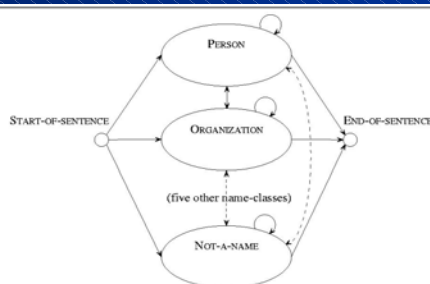


Figure 3.1 Pictorial representation of conceptual model. The subgraph of name-classes is complete, indicated here by the dashed arcs.

Identifinder HMM

- Generative model

1. Select a name-class NC , conditioning on the previous name-class and the previous word.
2. Generate the first word inside that name-class, conditioning on the current and previous name-classes.
3. Generate all subsequent words inside the current name-class, where each subsequent word is conditioned on its immediate predecessor (as per a standard bigram language model).

- Probability distributions needed:

$$- P(NC | NC_{-1}, w_{-1})$$

$$- P(\langle w, f \rangle_{first} | NC, NC_{-1})$$

$$- P(\langle w, f \rangle | \langle w, f \rangle_{-1}, NC_{-1})$$

IdentiFinder HMM

- Example: Mr. [John]_{E-PERSON} eats
Probability of the previous annotated sequence:

$\Pr(\text{NOT-A-NAME} \mid \text{START-OF-SENTENCE}, "+\text{end}+")^*$
 $\Pr("Mr." \mid \text{NOT-A-NAME}, \text{START-OF-SENTENCE})^*$
 $\Pr(+\text{end+} \mid "Mr.", \text{NOT-A-NAME})^*$
 $\Pr(\text{PERSON} \mid \text{NOT-A-NAME}, "Mr.")^*$
 $\Pr("Jones" \mid \text{PERSON}, \text{NOT-A-NAME})^*$
 $\Pr(+\text{end+} \mid "Jones", \text{PERSON})^*$
 $\Pr(\text{NOT-A-NAME} \mid \text{PERSON}, "Jones")^*$
 $\Pr("eats" \mid \text{NOT-A-NAME}, \text{PERSON})^*$
 $\Pr("." \mid "eats", \text{NOT-A-NAME})^*$
 $\Pr(+\text{end+} \mid ".", \text{NOT-A-NAME})^*$
 $\Pr(\text{END-OF-SENTENCE} \mid \text{NOT-A-NAME}, ".")^*$

IdentiFinder HMM

- Parameter Estimation
 - Maximum likelihood estimates
 - Incorporates smoothing and back-off to face sparseness
- Decoding
 - $\text{argmax}_{NC} (NC_1 \dots NC_n \mid w_1 \dots w_n)$
 - Viterbi algorithm (dynamic programming)

IdentiFinder: Results

Table 5.1 F-measure Scores. This table illustrates IdentiFinder's performance as compared to the best reported scores for each category.

	Language	Best Rules	IdentiFinder
Mixed Case	English (WSJ)	96.4	94.9
Upper Case	English (WSJ)	89	93.6
Speech Form	English (WSJ)	74	90.7
Mixed Case	Spanish	93	90

IdentiFinder: Results

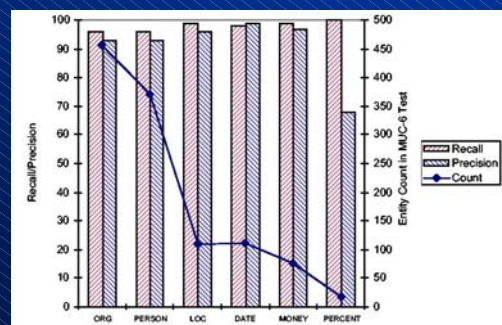


Figure 5.1 Detailed Performance Analysis of IdentiFinder on the MUC-6 Test. The left axis is percentage of recall/precision for the bars; the right axis is the raw entity count for the line graph. There are no occurrences of the TIME name-class in the MUC-6 test.

Low-cost Named Entity Classification for Catalan: Exploiting Multilingual Resources and Unlabeled Data

Lluís Márquez, Adrià de Gispert, Xavier Carreras and Lluís Padró
{lluism, agispert, carreras, padro}@talp.upc.es



MuNER Workshop - ACL 2003, Sapporo, July 12

Outline

- Introduction
- Data Resources
- Learning Algorithms
- Using only Catalan data
- Using Spanish resources
- Conclusions



MuNER workshop - ACL'03, Sapporo, 12 July 2003

10/6/2003

Outline

- **Introduction**
- Data Resources
- Learning Algorithms
- Using only Catalan data
- Using Spanish resources
- Conclusions



MuNER workshop - ACL'03, Sapporo, 12 July 2003

10/6/2003

Introduction

Introduction

ML-based Named Entity Classification for Catalan

- Building a low-cost ML-based system
- No available resources
- Use existing Spanish resources
- Take advantage of their similarities:
 - NE internal structure
 - NE context structure
 - Social and cultural environments



MuNER workshop - ACL'03, Sapporo, 12 July 2003

10/6/2003

Introduction

An Example

Spanish

"El presidente del Comité Olímpico Internacional, **José Antonio Samaranch**, se reunió el lunes en **Nueva York** con investigadores del FBI y del Departamento de Justicia."

Catalan

"El president del Comitè Olímpic Internacional, **Josep Antoni Samaranch**, es va reunir dilluns a **Nova York** amb investigadors del FBI i del Departament de Justícia."



MuNER workshop - ACL'03, Sapporo, 12 July 2003

10/6/2003

Outline

- Introduction
- **Data Resources**
- Learning Algorithms
- Using only Catalan data
- Using Spanish resources
- Conclusions



MuNER workshop - ACL'03, Sapporo, 12 July 2003

10/6/2003

Existing Catalan Resources

- Catalan corpus:
 - From the daily newspaper: "*El Periódico de Catalunya*"
 - Subset of ~2,200,000 words from year 2,000
 - No linguistic annotation
 - Manually annotated **train** (23,177 words, 1,232 NEs) and **test** sets (23,595 words, 1,338 NEs), with a labour cost of **10 person hours**
 - **Unlabelled** set (2,201,712 words, 75,038 NEs*)
 - * 91.5% accurate NER module (Carreras et al., 2003)



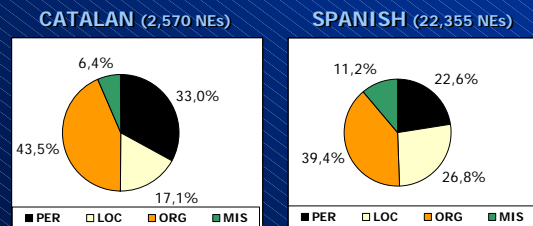
Existing Spanish Resources

- CoNLL-2002 corpus:
 - From the EFE Newswire Agency collection
 - ~365,000 words. ~26,000 NEs
 - Hand-tagged
 - Divided into **train** (~70%), **development** and **test** sets (~15% each)
- Four categories to be considered: PER, LOC, ORG and MIS



Distribution of categories

- Categories show the following distribution for hand-labelled data:



Feature codification

- Window-based
 - **Lexical**: *word forms*.
 - **Ortographic**: *capitalization, digits, hyphenation, etc.*
 - **Affixes**: *prefixes, suffixes (up to 4 letters)*
 - **Word-type patterns**: *functional, capitalized, lowercase, punctuation, quote, other.*
 - **Bag-of-words**: *word forms in window (no position)*
 - **Trigger words, and context patterns**
 - **Gazetteer**: *set of possible NE categories*
 - **Length of the NE**



Outline

- Introduction
- Data Resources
- **Learning Algorithms**
- Using only Catalan data
- Using Spanish resources
- Conclusions



Two learning approaches

- **Supervised**: Use existing Catalan hand-tagged data to train a classifier using the AdaBoost algorithm
- **Unsupervised**: Using the Greedy Agreement Algorithm to bootstrap from unlabelled Catalan data



Learning Algorithms

Supervised Learning

Use Catalan data to train Adaboost models

- AdaBoost constructs an ensemble of base (weak) classifiers, which are linearly combined for making real-valued predictions (Schapire & Singer, 1999)
- Base rules are implemented as small fixed-depth decision trees

MuNER workshop – ACL'03, Sapporo, 12 July 2003 10/6/2003

Learning Algorithms

Unsupervised Learning

Using the Greedy Agreement Algorithm (Abney, 2002)

- Two binary classifiers alternatively add new rules optimising agreement on unlabelled data

MuNER workshop – ACL'03, Sapporo, 12 July 2003 10/6/2003

Learning Algorithms

Unsupervised Learning

Using the Greedy Agreement Algorithm (Abney, 2002)

- Two binary classifiers alternatively add new rules optimising agreement on unlabelled data
- The classifiers are based on two “independent” views on the data
- Each classifier votes the contribution of several atomic rules → if suffix “lez” then PER
- A few atomic rules are manually introduced as seeds to start iterating

MuNER workshop – ACL'03, Sapporo, 12 July 2003 10/6/2003

Learning Algorithms

Unsupervised Learning

Using the Greedy Agreement Algorithm (Abney, 2002)

- Two **binary** classifiers alternatively add new rules optimising agreement on unlabelled data
- The classifiers are based on two **“independent” views** on the data
- Each classifier votes the contribution of several atomic rules → if suffix “lez” then PER
- A few atomic rules are manually introduced as **seeds** to start iterating

MuNER workshop – ACL'03, Sapporo, 12 July 2003 10/6/2003

Learning Algorithms

Unsupervised Learning

Using the Greedy Agreement Algorithm (Abney, 2002)

- Multiclass** by *one-vs-all* binarization (4 classifiers, combined). Strategies tested:
 - only valid when one positive and rest negative
 - combining all votes from all classifiers (best approach)
- Two alternatives for **view selection**
 - Local and contextual information (GAp)
 - Mixed approach (GAm) →
- Seed rules** selection
 - Blind selection of higher recall (>98%) rules for a validation set
 - Manual selection of generalising rules (best approach)

MuNER workshop – ACL'03, Sapporo, 12 July 2003 10/6/2003

GA Algorithm

The Greedy Agreement

(Abney, 2002)

- For each sample, each atomic rule related to each of its features votes for “negative” or “positive”
- Each rule can appear more than once (higher weight)

```

Input: seed rules F, G
loop
  for each atomic rule H
    G' = G + H
    evaluate cost of (F, G')
    keep lowest-cost G'
  if G' is worse than G, quit
  swap F, G'
  
```

MuNER workshop – ACL'03, Sapporo, 12 July 2003 10/6/2003

Outline

- Introduction
- Data Resources
- Learning Algorithms
- **Using only Catalan data**
- Using Spanish resources
- Conclusions

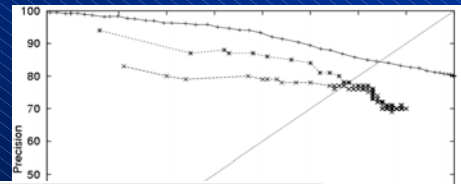


MulNER workshop – ACL'03, Sapporo, 12 July 2003

10/6/2003

Using only Catalan data

Supervised vs unsupervised



	LOC	ORG	PER	MIS	avg.
GA _p	14.66	83.64	93.88	0.00	66.66
GA _m	20.67	95.30	76.94	4.00	68.28
CA	61.65	86.84	91.67	40.00	79.83



MulNER workshop – ACL'03, Sapporo, 12 July 2003

10/6/2003

Using only Catalan data

Bootstrapping

M_0 =initial model
 T_L =initial labelled (training) corpus
for $i = 1 \dots N$ **do**
 - Classify NEs in $S_1 \dots S_i$ using model M_{i-1}
 - Select a **subset** S from previously classified samples $U_{j=1}^i S_j$
 - Learn a new model M_i using $T_L \cup S$ as training data
endfor
output model M_N



MulNER workshop – ACL'03, Sapporo, 12 July 2003

10/6/2003

Using only Catalan data

Bootstrapping

- Subset selection
 - All examples (bts1)
 - Most confident ones according to Adaboost (bts2)
 - Same as before plus most confident ones according to GA (bts3)

it.	CA _{bts1}	CA _{bts2}	CA _{bts3}
0	79.83	79.83	79.41
1	78.48	79.58	79.46
2	78.29	79.22	80.04
3	78.13	79.87	79.95
4	78.01	79.58	79.56
5	78.73	79.08	79.11
6	78.22	79.07	79.95
7	78.25	78.93	80.63
8	77.99	79.14	79.65
9	78.17	79.57	79.17
10	78.30	78.89	79.21



MulNER workshop – ACL'03, Sapporo, 12 July 2003

10/6/2003

Outline

- Introduction
- Data Resources
- Learning Algorithms
- Using only Catalan data
- **Using Spanish resources**
- Conclusions



MulNER workshop – ACL'03, Sapporo, 12 July 2003

10/6/2003

Using Spanish resources

Bilingual models

- Training a bilingual classifier from all labelled examples (Spanish and Catalan)
- Same set of features, except lexical features, which are translated
 $[w_j = \text{"calle"}] \Leftrightarrow [w_j = \text{"carrer"}]$
- Translation dictionary built automatically (InterNOSTRUM automatic translation + shallow supervision)
- Cross-linguistic features are used
 $[(w_j = \text{"calle"} \wedge \text{lang} = \text{es}) \vee (w_j = \text{"carrer"} \wedge \text{lang} = \text{ca})]$



MulNER workshop – ACL'03, Sapporo, 12 July 2003

10/6/2003

Bilingual models. Results

- Two best-performing bootstrapping strategies also applied

	LOC	ORG	PER	MIS	avg.
CA	61.65	86.84	91.67	40.00	79.83
CA _{bts3}	65.41	87.22	91.94	37.33	80.63
XL	69.17	88.16	92.76	45.33	82.63
XL _{bts2}	70.68	89.10	94.71	41.33	83.73

- Resultant model is useful for both languages
 - Spanish classification: from 87,1% to 86,9%



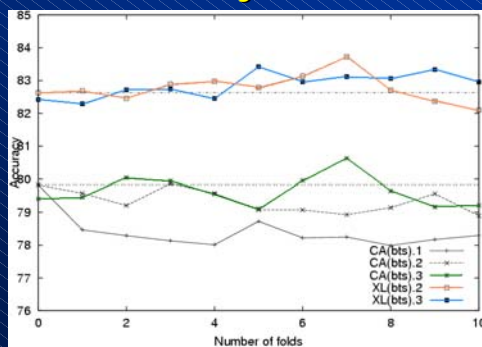
Bilingual models. Results

- Two best-performing bootstrapping strategies also applied (bts2 and bts3)
- More stable behaviour (smaller impact of Catalan examples in front of Spanish ones in the bilingual model)

it.	XL _{bts2}	XL _{bts3}
0	82.63	82.42
1	82.69	82.29
2	82.45	82.72
3	82.89	82.74
4	82.98	82.45
5	82.79	83.42
6	83.14	82.96
7	83.73	83.12
8	82.70	83.06
9	82.37	83.34
10	82.10	82.96



Summary Results



Outline

- Introduction
- Data Resources
- Learning Algorithms
- Using only Catalan data
- Using Spanish resources
- Conclusions**



Conclusions

- Given a small labelled data set, AdaBoost supervised learning clearly outperforms the fully supervised Greedy Agreement algorithm
- Supervised models trained with few annotated data do not easily profit from bootstrapping strategies
 - Examples from unsupervised models add complementary info
- Multilingual models improve accuracy significantly for the language with less annotated data, without a significant decrease in performance for language with more data



Low-cost Named Entity Classification for Catalan: Exploiting Multilingual Resources and Unlabeled Data

Lluís Màrquez, Adrià de Gispert, Xavier Carreras and Lluís Padró

{lluism, agispert, carreras, padro}@talp.upc.es



MulNER Workshop - ACL 2003, Sapporo, July 12

Wide-Coverage Spanish Named Entity Extraction

Xavier Carreras, Lluís Màrquez & Lluís Padró



TALP Research Center

LSI, UPC
Technical University of Catalonia

IBERAMIA'02

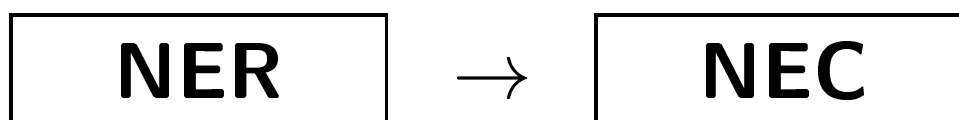
Wide-Coverage Spanish Named Entity Extraction

Named Entity Extraction

- The problem:

“Según informó el (**Departamento**)_{ORG} que dirige el consejero (**Inaxio Oliveri**)_{PER} , los representantes de la (**Consejería**)_{ORG} y de las universidades del (**País Vasco**)_{LOC}, (**Deusto**)_{ORG} y (**Mondragón**)_{ORG} estudiaron los nuevos retos de estos centros educativos.”

- Useful for: Information extraction, improving linguistic processors, automatic summarization, document linking/clustering, etc.
- Our Approach: Named Entity Extraction as two separate modules:



- **NER** = sequence tagging (e.g., IOB tagging)
- **NEC** = classification task

Named Entity Extraction

- All (local) decisions resolved with Machine Learning–based classifiers
- Learning Algorithm: **AdaBoost**
 - Combination of many *weak* classifiers (hypotheses or rules) into a *strong* classifier
 - Binary classification, binary features
 - Real-valued AdaBoost with confidence-rated predictions (Schapire & Singer 99)
 - Combined classifier: $f(x) = \sum_{t=1}^T \alpha_t h_t(x)$
 - Weak Rules (h_t): Decision Trees of fixed depth
 - Good performance/efficiency in NLP domains (Abney et al., 99; Schapire & Singer, 00; Escudero et al., 00) (Carreras & Màrquez, 01; Carreras et al., 02a; 02b; etc.)

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

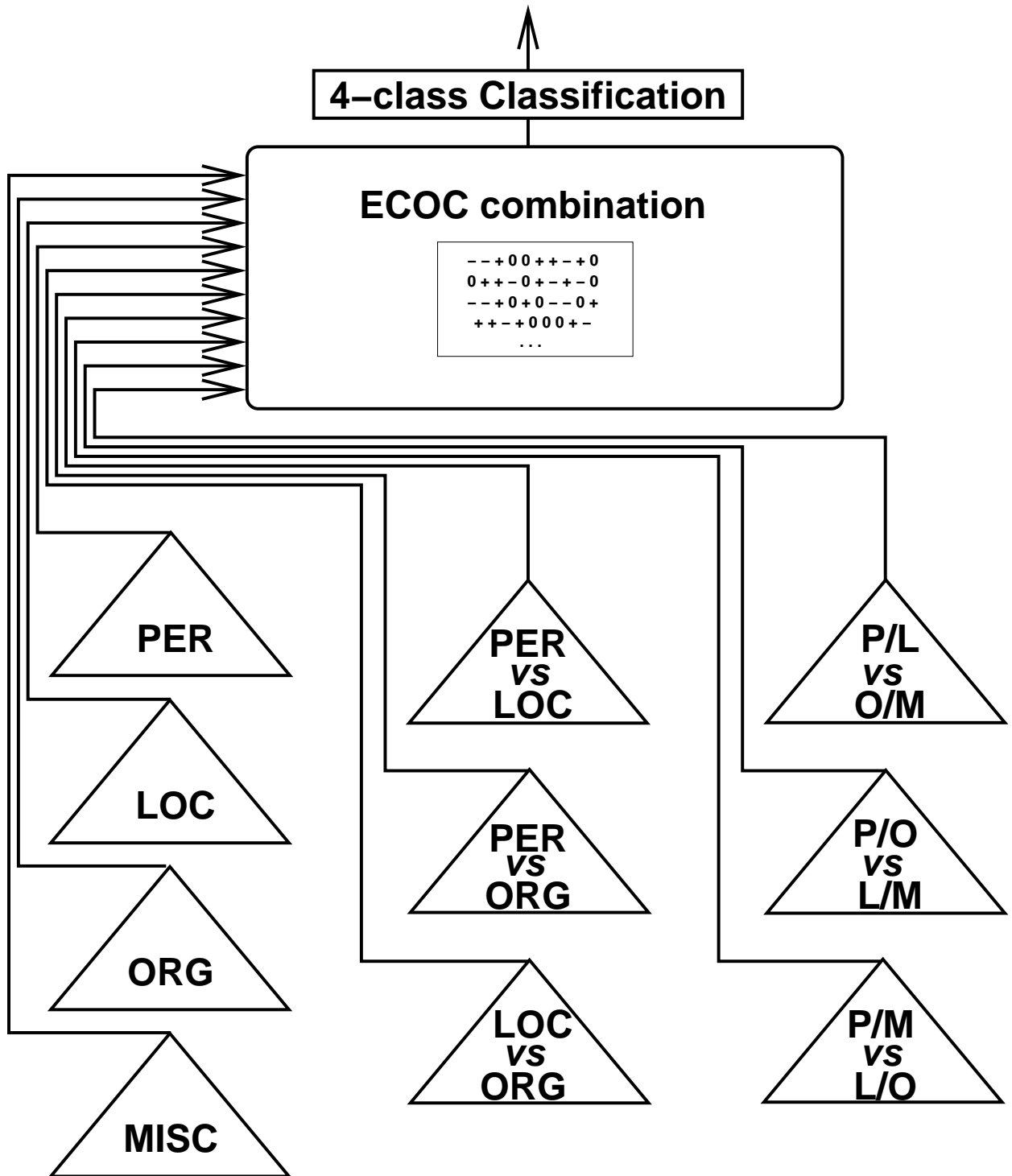
- greedy Open-Close, plus Inside checking:

	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

NEC Module



Features

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

<i>initial-caps</i>	<i>all-caps</i>	<i>all-digits</i>
<i>roman-number</i>	<i>contains-dots</i>	<i>contains-hyphen</i>
<i>acronym</i>	<i>lonely-initial</i>	<i>punctuation-mark</i>
<i>single-char</i>	<i>functional-word</i>	<i>URL</i>

- 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 (BIO or Open/Close tags)

Evaluation

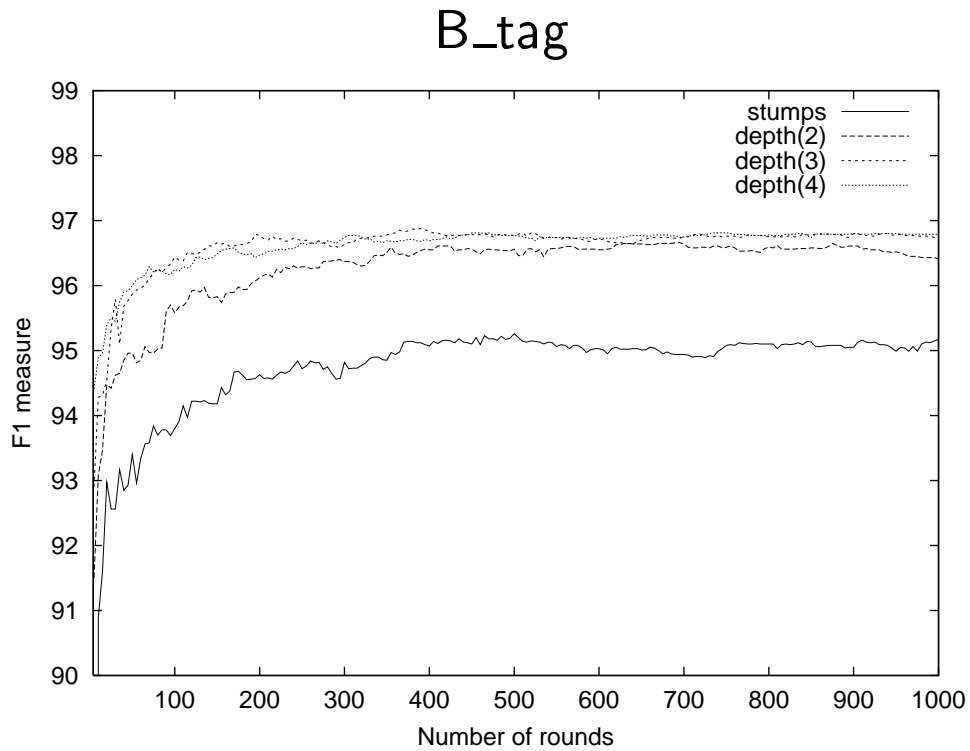
- The(*Agencia*) EFE Spanish corpus:
 - Collection of over 3,000 news agency articles issued during year 2000
 - 802,729 words, which contain about 86,000 hand tagged named entities
- Experimental setting:
 - Test subset: $\sim 65,000$ words (4,820 NE)
 - NEC training set: $\sim 700,000$ words (61,266 NE)
 - NER training set: $\sim 100,000$ words (8,126 NE)
 - Only features occurring more than 2 times
 - Learning parameters: *number of rounds, depth of the base classifiers* (from stumps to decision trees of depth 4)
 - Evaluation measures: *precision, recall, F_1 , accuracy*

Learning Problems: Sizes

NER	#Exs.	#Feat.	#Pos.examples
open	91,625	19,215	8,126 (8.87%)
close	8,802	10,795	4,820 (54.76%)
I	97,333	20,526	5,708 (5.86%)
O	97,333	20,526	83,499 (85.79%)
B	97,333	20,526	8,126 (8.35%)

NEC	#Exs.	#Feat.	#Pos.examples
PERSON	61,266	22,465	12,976 (21.18%)
LOCATION	61,266	22,465	14,729 (24.04%)
ORGANIZ	61,266	22,465	22,947 (34.46%)
OTHER	61,266	22,465	10,614 (17.32%)

Results on the NER Task (1)



Method	P	R	F_1
MACO+	89.94%	87.51%	88.71%
OpenClose	92.42%	91.54%	91.97%
BIO	92.66%	91.99%	92.33%
OpenClose+I	92.60%	92.14%	92.37%

Results on the NER Task (2)

Subset	#NE	<i>P</i>	<i>R</i>	<i>F</i>₁
length=1	2,807	94.64%	95.65%	95.15%
length=2	1,005	94.01%	93.73%	93.87%
length=3	495	91.65%	88.69%	90.14%
length=4	237	84.81%	84.81%	84.81%
length=5	89	77.27%	76.40%	76.84%
length=6	74	81.94%	79.73%	80.82%
length=7	22	60.00%	54.55%	57.14%
length=8	22	88.24%	68.18%	76.92%
length=9	11	80.00%	72.73%	76.19%
length=10	3	50.00%	33.33%	40.00%
uppercase	4,637	92.81%	93.25%	93.03%
lowercase	183	85.40%	63.93%	73.13%
TOTAL	4,820	92.60%	92.14%	92.37%

Results on the NEC Task

Method	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Acc</i>
Most freq	39.78%	39.78%	39.78%	39.78%
basic	90.19%	84.44%	87.22%	87.51%
+tw	90.11%	84.77%	87.36%	88.17%
+gaz	90.25%	85.31%	87.71%	88.60%
+tw+gaz	90.61%	85.23%	87.84%	88.73%

Method	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Acc</i>
Most freq	37.47%	37.47%	37.47%	37.47%
basic	83.84%	79.39%	81.55%	81.85%
+tw	85.08%	79.30%	82.09%	82.15%
+gaz	85.05%	79.57%	82.22%	82.15%
+tw+gaz	85.32%	79.80%	82.47%	82.31%

Comparison to other Systems

- CoNLL'02 Shared Task on NERC
 - Organized by the ACL's SIG on NLL (Sep.02)
 - Spanish and Dutch datasets publicly available
 - 12 participants using ML-based systems
 - Best results in both languages (~ 2 points F_1)
 - Why?
 - * Appropriateness of the learning algorithm
 - * Design of the learning attributes
 - * Combination of classifiers

Conclusions and Further Work

- Conclusions:
 - NERC system for Spanish based on robust Machine Learning techniques
 - Large set of simple features requiring no complex linguistic processing
 - Fairly good performance (validated in the CoNLL'02 competition framework)
 - Gazetteers and trigger words allow to slightly improve performance
- Current and future work:
 - Classification algorithms other than AdaBoost (e.g., SVMs)
 - Use of global inference schemes
 - Multi-label AdaBoost algorithms
 - Reusing of the Spanish models to develop a system for Catalan (work submitted to EACL)