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

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
Named Entity Recognition

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.
Named Entity Recognition

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–)
- IdentiFinder™
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 Methods for NER

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
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 → NEC](#)

  - **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:**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
<th>$w_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>2</td>
<td>-1</td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>output</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
</tbody>
</table>

- **greedy Open-Close, plus Inside checking:**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
<th>$w_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>2</td>
<td>3</td>
<td></td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>3</td>
<td></td>
<td>2</td>
<td></td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close</td>
<td>-1</td>
<td>2</td>
<td>3</td>
<td>-2</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>output</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
</tbody>
</table>

- **global Open-Close:**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
<th>$w_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>x</td>
<td>O</td>
<td>x</td>
<td>x</td>
<td>O</td>
<td>O</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Close</td>
<td>x</td>
<td>x</td>
<td>C</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>C</td>
<td>x</td>
</tr>
<tr>
<td>g1 1</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>O</td>
</tr>
<tr>
<td>g1 2</td>
<td>x</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>I</td>
<td>O</td>
</tr>
<tr>
<td>g1 3</td>
<td>x</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
<tr>
<td>output</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
</tbody>
</table>

Wide–Coverage Spanish Named Entity Extraction
NEC Module

4-class Classification

ECOC combination

- - + 0 + + - 0
0 + - 0 + - + 0
- - + 0 + - 0 -
+ + + + 0 0 + -
...

Wide-Coverage Spanish Named Entity Extraction
Features

- Sliding Window, codifying ±3 words.

- Primitive Features in the window:
  - Word form
  - PoS (when available)
  - Ortographic Features
    - \textit{initial-caps} \quad \textit{all-caps} \quad \textit{all-digits}
    - \textit{roman-number} \quad \textit{contains-dots} \quad \textit{contains-hyphen}
    - \textit{acronym} \quad \textit{lone-\text{\textemdash\textendash}initial} \quad \textit{punctuation-mark}
    - \textit{single-char} \quad \textit{functional-word} \quad \textit{URL}

- Word-Type Patterns:
  - \textit{functional} \quad \textit{uppercase} \quad \textit{lowercase} \quad \textit{punctuation} \quad \textit{quote} \quad \textit{other}

- Bag-of-Words (±5 window)
- Trigger Words (class of); also patterns
- Gazetteer Features (class of)
- Left Predictions (BIO or Open/Close tags)
Weak Rules

From Open classifier

\[
\begin{align*}
o_0 & : \text{Caps} -- t_1 & : \text{Out} -- w_2 & : \text{NA} -- -0.430 \\
| & & & \\
| & & & -- 1.058 \\
| & & & \\
| & & & -- o_3 & : \text{allCaps} -- -0.859 \\
| & & & \\
| & & & -- -2.203 \\
| & & & \\
-- w_0 & : " -- o_1 & : \text{Caps} -- -0.092 \\
| & & & \\
| & & & -- -2.614 \\
| & & & \\
| & & & -- o_2 & : \text{Fun} -- -2.799 \\
| & & & \\
| & & & -- -4.318
\end{align*}
\]
Weak Rules (2)

From Close classifier

```
or1:Cap -- or3:allCap -- oB:allCap -- 0.473
|                          |                           |
|                          |                           -- -1.285
|                          |
|                          -- or2:Cap -- -1.366
|                          |
|                          -- -2.413
|                          -- or1:Fun -- or2:Cap -- -0.976
|                          |
|                          -- 0.252
|                          -- or1:allCap -- -2.191
|                          -- 1.741
```

Wide-Coverage Spanish Named Entity Extraction
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: ~65,000 words (4,820 NE)
  - NEC training set: ~700,000 words (61,266 NE)
  - NER training set: ~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

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

<table>
<thead>
<tr>
<th>NEC</th>
<th>#Exs.</th>
<th>#Feat.</th>
<th>#Pos. examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>61,266</td>
<td>22,465</td>
<td>12,976 (21.18%)</td>
</tr>
<tr>
<td>LOCATION</td>
<td>61,266</td>
<td>22,465</td>
<td>14,729 (24.04%)</td>
</tr>
<tr>
<td>ORGANIZ</td>
<td>61,266</td>
<td>22,465</td>
<td>22,947 (34.46%)</td>
</tr>
<tr>
<td>OTHER</td>
<td>61,266</td>
<td>22,465</td>
<td>10,614 (17.32%)</td>
</tr>
</tbody>
</table>
Results on the NER Task (1)

### Method

<table>
<thead>
<tr>
<th>Method</th>
<th>$P$</th>
<th>$R$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maco+</td>
<td>89.94%</td>
<td>87.51%</td>
<td>88.71%</td>
</tr>
<tr>
<td>OpenClose</td>
<td>92.42%</td>
<td>91.54%</td>
<td>91.97%</td>
</tr>
<tr>
<td>BIO</td>
<td>92.66%</td>
<td>91.99%</td>
<td>92.33%</td>
</tr>
<tr>
<td>OpenClose+I</td>
<td>92.60%</td>
<td>92.14%</td>
<td>92.37%</td>
</tr>
</tbody>
</table>

Wide-Coverage Spanish Named Entity Extraction
## Results on the NER Task (2)

<table>
<thead>
<tr>
<th>Subset</th>
<th>#NE</th>
<th>$P$</th>
<th>$R$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>length=1</td>
<td>2,807</td>
<td>94.64%</td>
<td>95.65%</td>
<td>95.15%</td>
</tr>
<tr>
<td>length=2</td>
<td>1,005</td>
<td>94.01%</td>
<td>93.73%</td>
<td>93.87%</td>
</tr>
<tr>
<td>length=3</td>
<td>495</td>
<td>91.65%</td>
<td>88.69%</td>
<td>90.14%</td>
</tr>
<tr>
<td>length=4</td>
<td>237</td>
<td>84.81%</td>
<td>84.81%</td>
<td>84.81%</td>
</tr>
<tr>
<td>length=5</td>
<td>89</td>
<td>77.27%</td>
<td>76.40%</td>
<td>76.84%</td>
</tr>
<tr>
<td>length=6</td>
<td>74</td>
<td>81.94%</td>
<td>79.73%</td>
<td>80.82%</td>
</tr>
<tr>
<td>length=7</td>
<td>22</td>
<td>60.00%</td>
<td>54.55%</td>
<td>57.14%</td>
</tr>
<tr>
<td>length=8</td>
<td>22</td>
<td>88.24%</td>
<td>68.18%</td>
<td>76.92%</td>
</tr>
<tr>
<td>length=9</td>
<td>11</td>
<td>80.00%</td>
<td>72.73%</td>
<td>76.19%</td>
</tr>
<tr>
<td>length=10</td>
<td>3</td>
<td>50.00%</td>
<td>33.33%</td>
<td>40.00%</td>
</tr>
<tr>
<td>uppercase</td>
<td>4,637</td>
<td>92.81%</td>
<td>93.25%</td>
<td>93.03%</td>
</tr>
<tr>
<td>lowercase</td>
<td>183</td>
<td>85.40%</td>
<td>63.93%</td>
<td>73.13%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>4,820</td>
<td>92.60%</td>
<td>92.14%</td>
<td>92.37%</td>
</tr>
</tbody>
</table>
## Results on the NEC Task

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

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most freq</td>
<td>37.47%</td>
<td>37.47%</td>
<td>37.47%</td>
<td>37.47%</td>
</tr>
<tr>
<td>basic</td>
<td>83.84%</td>
<td>79.39%</td>
<td>81.55%</td>
<td>81.85%</td>
</tr>
<tr>
<td>+tw</td>
<td>85.08%</td>
<td>79.30%</td>
<td>82.09%</td>
<td>82.15%</td>
</tr>
<tr>
<td>+gaz</td>
<td>85.05%</td>
<td>79.57%</td>
<td>82.22%</td>
<td>82.15%</td>
</tr>
<tr>
<td>+tw+gaz</td>
<td><strong>85.32%</strong></td>
<td><strong>79.80%</strong></td>
<td><strong>82.47%</strong></td>
<td><strong>82.31%</strong></td>
</tr>
</tbody>
</table>
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

View of Solving NLP Problems

Weaknesses of Pipeline Model
- Propagation of errors
- Bi-Directional interactions between stages
  - Occasionally, later stage problems are easier.
  - Upstream mistakes will not be corrected.

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.

Ideal Inference

Inference Procedure
- Inference with classifiers is not a new idea.
  - On sequential constraint structure:
    - HMM, PMM, CRF (Lafferty et al.), CSCL (Punyakarnik&Bontemps)
  - 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
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

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):

\[
\begin{align*}
P_1: & \quad 0.6 \quad 0.4 \\
P_2: & \quad 0.7 \quad 0.3 \\
P_3: & \quad 0.8 \quad 0.2 \\
P_4: & \quad 0.9 \quad 0.1 \\
\end{align*}
\]

NP: Noun Phrase
∅: Null (not a phrase)

Cost = 0.6 + 0.3 + 0.2 + 0.9 = 2.0

Phrase Identification Problems

Several NLP problems attempt to detect and classify phrases in sentences.

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

LP Formulation – Linear Cost

- Indicator variables
  
  \[
  x_{[P1=NP]}, x_{[P1=∅]}, \ldots, x_{[P4=NP]}, x_{[P4=∅]} \in \{0, 1\}
  \]

**Total Cost**

\[
\begin{align*}
\text{Cost} &= c_{[P1=NP]} \cdot x_{[P1=NP]} + c_{[P1=∅]} \cdot x_{[P1=∅]} + \ldots + c_{[P4=∅]} \cdot x_{[P4=∅]} \\
&= 0.6 \cdot x_{[P1=NP]} + 0.4 \cdot x_{[P1=∅]} + \ldots + 0.1 \cdot x_{[P4=∅]}
\end{align*}
\]

LP Formulation – Linear Constraints

Subject to:

**Non-overlapping Constraints**:

\[
\begin{align*}
x_{[P1=∅]} + x_{[P3=∅]} & \geq 1 \\
x_{[P2=∅]} + x_{[P3=∅]} + x_{[P4=∅]} & \geq 2
\end{align*}
\]

LP Formulation

\[
\begin{align*}
\text{max} & \quad \sum_{i \in \text{Phrases}, k \in \text{Classes}} c_{i,k} \cdot x_{i,k} \\
\text{subject to:} & \\
& \quad x_{i,k} \in \{0, 1\} \quad \forall i \in \text{Phrases}, k \in \text{Classes} \\
& \quad \sum_{k \in \text{Classes}} x_{i,k} = 1 \quad \forall i \in \text{Phrases} \\
& \quad \sum_{i=1}^{m} x_{i,j,k} = m - 1 \quad \forall P_j, 1 \leq i \leq m, \text{ that overlap}
\end{align*}
\]

Generate one integer linear program per sentence.

Entity/Relation Recognition

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

Identify:

- Kill (X, Y)

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

1. Entity boundaries are given.
2. The relation between each pair of entities is represented by a relation variable; most of them are null.
3. The goal is to assign labels to these $E$ and $R$ variables.

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

LP Formulation – Indicator Variables

For each variable:
- $x_{[E_1 = \text{per}]}$, $x_{[E_1 = \text{loc}]}$, ..., $x_{[R_{12} = \emptyset]}$, ..., $x_{[R_{12} = \emptyset]} \in \{0,1\}$

For each pair of variables on an edge:
- $x_{[R_{12} = \text{kill}, E_1 = \text{per}]}$, $x_{[R_{12} = \text{kill}, E_1 = \text{loc}]}$, ..., $x_{[R_{12} = \emptyset, E_1 = \text{loc}]}$, ...

LP Formulation – Cost Function

- Assignment cost
- Constraint cost

Total cost = Assignment cost + Constraint cost

LP Formulation – Linear Constraints

Subject to:
- Node—Edge Consistency Constraints

Experiments – Data

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

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

Decision-time Constraint

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

$$X_{[R12=\text{kill}]} \cdot X_{[R21=\text{kill}]} \cdot X_{[R13=\text{kill}]} + ... \geq 1$$

Suppose we know the given sentence has the "kill" relation

Computational Issues

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

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-Obj constraints
  - Able to handle non-sequential constraint structure
    - E/R case has demonstrated this property

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

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]
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).

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, all-digits, all-dots, etc.
• Model M3 - adds part of speech (POS) attributes (using TnT)
• Model M4 - adds basic syntactic phrases (chunks) attributes (using YamCha)
• 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

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>M7</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC</td>
<td>81.13</td>
<td>83.50</td>
<td>83.02</td>
<td>83.34</td>
<td>82.71</td>
<td>82.18</td>
<td>83.11</td>
</tr>
<tr>
<td>MISC</td>
<td>56.14</td>
<td>65.57</td>
<td>65.39</td>
<td>65.24</td>
<td>65.48</td>
<td>66.49</td>
<td>66.36</td>
</tr>
<tr>
<td>MONEY</td>
<td>76.60</td>
<td>74.61</td>
<td>75.00</td>
<td>75.79</td>
<td>81.22</td>
<td>88.39</td>
<td>82.59</td>
</tr>
<tr>
<td>ORG</td>
<td>62.63</td>
<td>63.33</td>
<td>64.23</td>
<td>64.91</td>
<td>64.25</td>
<td>64.16</td>
<td>64.83</td>
</tr>
<tr>
<td>TIME</td>
<td>62.48</td>
<td>74.27</td>
<td>72.73</td>
<td>70.76</td>
<td>69.22</td>
<td>70.30</td>
<td>77.72</td>
</tr>
<tr>
<td>Overall</td>
<td>75.12</td>
<td>73.96</td>
<td>73.83</td>
<td>74.04</td>
<td>74.32</td>
<td>74.24</td>
<td>70.78</td>
</tr>
</tbody>
</table>

Best model = M1 + M2 + M6 + M7

Experimental Results

(2) F measure on SWB transcripts without case information

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>M7</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC</td>
<td>80.86</td>
<td>81.29</td>
<td>80.15</td>
<td>80.27</td>
<td>79.07</td>
<td>79.06</td>
<td>79.90</td>
</tr>
<tr>
<td>MISC</td>
<td>53.89</td>
<td>54.62</td>
<td>53.42</td>
<td>53.15</td>
<td>48.81</td>
<td>49.37</td>
<td>50.68</td>
</tr>
<tr>
<td>MONEY</td>
<td>78.31</td>
<td>78.72</td>
<td>77.49</td>
<td>76.84</td>
<td>79.59</td>
<td>79.00</td>
<td>79.41</td>
</tr>
<tr>
<td>ORG</td>
<td>61.79</td>
<td>62.68</td>
<td>62.68</td>
<td>63.19</td>
<td>58.71</td>
<td>59.85</td>
<td>58.10</td>
</tr>
<tr>
<td>PER</td>
<td>62.72</td>
<td>65.07</td>
<td>67.80</td>
<td>69.32</td>
<td>59.48</td>
<td>62.51</td>
<td>69.80</td>
</tr>
<tr>
<td>TIME</td>
<td>78.13</td>
<td>75.53</td>
<td>75.50</td>
<td>75.80</td>
<td>74.75</td>
<td>75.89</td>
<td>74.45</td>
</tr>
<tr>
<td>Overall</td>
<td>70.09</td>
<td>70.56</td>
<td>70.08</td>
<td>69.67</td>
<td>67.55</td>
<td>68.13</td>
<td>69.22</td>
</tr>
</tbody>
</table>

Best model = M1 + M2 + M6 + M7

Experimental Results

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

<table>
<thead>
<tr>
<th></th>
<th>Written Text (CoNLL)</th>
<th>Speech (SWB with case)</th>
<th>Speech (SWB without case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC</td>
<td>71.55</td>
<td>75.50</td>
<td>74.45</td>
</tr>
<tr>
<td>MISC</td>
<td>75.50</td>
<td>80.52</td>
<td>80.52</td>
</tr>
<tr>
<td>ORG</td>
<td>75.50</td>
<td>80.52</td>
<td>80.52</td>
</tr>
<tr>
<td>PER</td>
<td>75.50</td>
<td>80.52</td>
<td>80.52</td>
</tr>
<tr>
<td>TIME</td>
<td>75.50</td>
<td>80.52</td>
<td>80.52</td>
</tr>
<tr>
<td>Overall</td>
<td>71.55</td>
<td>75.50</td>
<td>74.45</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>LOC</th>
<th>MISC</th>
<th>ORG</th>
<th>PER</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-12.12</td>
<td>-43.45</td>
<td>-30.66</td>
<td>22.19</td>
<td>-24.68</td>
</tr>
</tbody>
</table>

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

Outline

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

Outline

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

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."
Introduction

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

An Example

Spanish

Comité Olímpico Internacional

Catalan

Comitè Olímpic Internacional

Outline

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

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)

Spanish NER Module

• 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

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:
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:

\[
\begin{align*}
\text{Caps}(w_0) & = \text{"de"} \\
\text{Caps}(w+1) & = \text{"del"} \\
\text{Caps}(w, j) & = \text{"I"} \\
\text{Tag}(w, j) & = \text{"Y"} \\
\end{align*}
\]

- \( w_0 \) = "de"
- \( w_0 \) = "del"
- \( w_0 \) = "I"
- \( w_0 \) = "Y"

Existing Catalan Resources
- Catalan corpus:
  - From the daily newspaper: "El Periódico de Catalunya"
  - Subset of ~2,200,000 words from the beginning of year 2000
  - No linguistic annotation

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.

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

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

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

Approaches
- Classcial: 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.
**Approaches: Translation**

- Models acquired on existing Spanish training data: 264,715 words containing 18,797 NEs
- Translation of lexical features (<5,000), e.g.: \([w_{-1}="calle"] \Rightarrow [w_{-1}="carrer"]\)
- Translation labour cost:
  - 10 person hours ("common-sense" hand translation)
  - 0.5 person hours (InterNOSTRUM automatic translation + shallow supervision)

**Approaches: X-Ling**

- Compilation of a dictionary of translation pairs
- Makes use of bilingual features, e.g.:
  \([(w_{-1}="calle\land\text{lang}=es}) \lor (w_{-1}="carrer\land\text{lang}=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**

<table>
<thead>
<tr>
<th></th>
<th>Feature trans.</th>
<th>en test</th>
<th>cn test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Straight</td>
<td></td>
<td>89.31</td>
<td>88.07</td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td>92.81</td>
<td>92.89</td>
</tr>
<tr>
<td>X-Ling (es)</td>
<td></td>
<td>92.25</td>
<td>92.64</td>
</tr>
<tr>
<td>X-Ling (mix)</td>
<td></td>
<td>92.27</td>
<td>92.53</td>
</tr>
</tbody>
</table>

**Evaluation**

<table>
<thead>
<tr>
<th></th>
<th>Feature trans.</th>
<th>en test</th>
<th>cn test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Straight</td>
<td></td>
<td>89.31</td>
<td>88.07</td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td>92.81</td>
<td>92.89</td>
</tr>
<tr>
<td>X-Ling (es)</td>
<td></td>
<td>92.25</td>
<td>92.64</td>
</tr>
<tr>
<td>X-Ling (mix)</td>
<td></td>
<td>92.27</td>
<td>92.53</td>
</tr>
</tbody>
</table>

**Evaluation**

<table>
<thead>
<tr>
<th></th>
<th>Feature trans.</th>
<th>en test</th>
<th>cn test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Straight</td>
<td></td>
<td>89.31</td>
<td>88.07</td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td>92.81</td>
<td>92.89</td>
</tr>
<tr>
<td>X-Ling (es)</td>
<td></td>
<td>92.25</td>
<td>92.64</td>
</tr>
<tr>
<td>X-Ling (mix)</td>
<td></td>
<td>92.27</td>
<td>92.53</td>
</tr>
</tbody>
</table>
Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature</th>
<th>en tran.</th>
<th>en test</th>
<th>en tran.</th>
<th>en test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>en</td>
<td>—</td>
<td>90.98</td>
<td>87.44</td>
<td>89.18</td>
</tr>
<tr>
<td>Straight</td>
<td>en</td>
<td>89.31</td>
<td>88.03</td>
<td>87.87</td>
<td>92.50</td>
</tr>
<tr>
<td>Translation</td>
<td>en</td>
<td>max.</td>
<td>92.81</td>
<td>92.49</td>
<td>92.00</td>
</tr>
<tr>
<td>X-Ling (es)</td>
<td>en</td>
<td>max.</td>
<td>92.25</td>
<td>92.64</td>
<td>92.48</td>
</tr>
<tr>
<td>X-Ling (mix)</td>
<td>en</td>
<td>max.</td>
<td>92.57</td>
<td>92.93</td>
<td>92.46</td>
</tr>
</tbody>
</table>

- Obtained models may be used in a bootstrapping process provided unannotated Catalan data is available.
- Unannotated corpus: 2.2 Mwords, divided into $S_1$…$S_N$ disjoint subsets of 1,000 sentences each.
- Iteratively annotate a growing number of 3 subsets and retrain using all annotated corpus.

Outline

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

Bootstrapping Procedure

$M_0$ = initial model  
$T_L$ = initial labelled (training) corpus

for $i = 1$…$N$ do
  - Identify NEs in $S_i$…$S_N$ using model $M_{i-1}$
  - Learn a new model $M_i$ using $T_L$ $U$ $U_{j=1}^{i-1}$ $S_j$ as training data
endfor

output model $M_N$
**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.

**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 NER system for Catalan.
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: \( \{o_1, \ldots, o_n\} \)
- sequence of states: \( \{s_1, \ldots, s_n\} \)

\[
\arg \max_{a_{1:n}} P(s_1, \ldots, s_n \mid o_1, \ldots, o_n) \approx \arg \max_{a_{1:n}} \prod_{k=1}^n P(s_k \mid s_{k-1}, s_{k-1}) \cdot P(o_k \mid s_k)
\]

- emission probabilities: \( P(o_k \mid s_k) \)
- transition probabilities: \( P(s_k \mid s_{k-1}, s_{k-1}) \)
- initial state probabilities: \( P(s_1) \)

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

IdentiFinder HMM
- Generative model
- Probability distributions needed:
  - \( P(\text{NC} \mid \text{NC}_{-1}, w_{-1}) \)
  - \( P(<w,f>_{\text{first}} \mid \text{NC}, \text{NC}_{-1}) \)
  - \( P(<w,f> \mid <w,f>_{-1}, \text{NC}_{-1}) \)

Figure 3.1: Pictorial representation of conceptual model. The subgraph of name-classes is accepted, indicated here by the hatched area.
IdentiFinder HMM

- Example: Mr. [John]E-PERSON eats
  Probability of the previous annotated sequence:
  
  \[ P(T(\text{NOT-A-NAME} | \text{START-OF-SENTENCE}, \text{+end+}) \) \times P(T(\text{Mr.} | \text{NOT-A-NAME}, \text{START-OF-SENTENCE}) \) \times P(T(\text{Mr.} | \text{NOT-A-NAME}) \) \times P(T(\text{Jones} | \text{PERSON}, \text{NOT-A-NAME}) \) \times P(T(\text{eats} | \text{NOT-A-NAME}, \text{PERSON}) \) \times P(T(\text{\text{"\}} | \text{NOT-A-NAME}, \text{\"\}) \) \times P(T(\text{END-OF-SENTENCE} | \text{NOT-A-NAME}, \text{\"\}) \) \]

- Parameter Estimation
  - Maximum likelihood estimates
  - Incorporates smoothing and back-off to face sparseness

- Decoding
  - \( \text{argmax}_{NC} (NC_1...NC_n|w_1...w_n) \)
  - Viterbi algorithm (dynamic programming)

IdentiFinder: Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Language</th>
<th>Best Rules</th>
<th>IdentiFinder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed Case</td>
<td>English (WSJ)</td>
<td>96.4</td>
<td>94.9</td>
</tr>
<tr>
<td>Upper Case</td>
<td>English (WSJ)</td>
<td>89</td>
<td>93.6</td>
</tr>
<tr>
<td>Speech Form</td>
<td>English (WSJ)</td>
<td>74</td>
<td>90.7</td>
</tr>
<tr>
<td>Mixed Case</td>
<td>Spanish</td>
<td>93</td>
<td>90</td>
</tr>
</tbody>
</table>

Figure 5.1: Detailed Performance Analysis of IdentiFinder on the MSCat-Eval. The left axis represents the percentage of correct solutions for the test data, while the right axis shows the number of correct solutions for each category. The bars illustrate the distribution of results across different language categories.
Low-cost Named Entity Classification for Catalan: Exploiting Multilingual Resources and Unlabeled Data

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

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

Outline

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

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

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.”
Existing Catalan Resources

- Catalan corpus:
  - From the daily newspaper: "El Periódico de Catalunya"
  - Subset of ~2,200,000 words from year 2000
  - 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:
  - CATALAN (2,570 NEs) SPANISH (22,355 NEs)

Feature codification

- 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.

- 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

**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

**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

**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)

**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 P, Q
Loop
  for each atomic rule H
    C' = Q + H
    evaluate cost of (F,C')
    keep lowest-cost C'
    if C' is worse than G, quit swap F, G'
Outline

• Introduction
• Data Resources
• Learning Algorithms

  Using only Catalan data
  Using Spanish resources
• Conclusions

Supervised vs unsupervised

Using only Catalan data

Bootstrapping

$M_0$= initial model
$T_L$= initial labelled (training) corpus

for $i = 1 ... N$

- Classify NEs in $S_1 ... S_i$ using model $M_{i-1}$
- Select a subset $S$ from previously classified samples $U_{i-j} \subseteq S_i$
- Learn a new model $M_i$ using $T_L U S$ as training data
endfor

output model $M_N$

Using only Catalan data

Bilingual models

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

Using Spanish resources

- Resultant model is useful for both languages
  - Spanish classification: from 87.1% to 86.9%

Summary Results

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

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

Outline

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

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

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

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

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
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:

\[ \text{NER} \rightarrow \text{NEC} \]

- 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:**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
<th>$w_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>2</td>
<td>-1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>3</td>
</tr>
<tr>
<td>output</td>
<td>0</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
</tbody>
</table>

- **greedy Open-Close, plus Inside checking:**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
<th>$w_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>2</td>
<td>3</td>
<td></td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>3</td>
<td></td>
<td>2</td>
<td></td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close</td>
<td>-1</td>
<td>2</td>
<td>3</td>
<td>-2</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>output</td>
<td>0</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
</tbody>
</table>

- **global Open-Close:**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
<th>$w_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>x</td>
<td>O</td>
<td>x</td>
<td>x</td>
<td>O</td>
<td>O</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Close</td>
<td>x</td>
<td>x</td>
<td>C</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>C</td>
<td>x</td>
</tr>
<tr>
<td>gl 1</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>O</td>
</tr>
<tr>
<td>gl 2</td>
<td>x</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>I</td>
<td>O</td>
</tr>
<tr>
<td>gl 3</td>
<td>x</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
<tr>
<td>output</td>
<td>0</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>B</td>
<td>B</td>
<td>I</td>
<td>O</td>
</tr>
</tbody>
</table>

Wide-Coverage Spanish Named Entity Extraction
NEC Module

4-class Classification

ECOC combination

- - + 0 0 + - + 0
0 + + 0 - + - 0
- - + 0 + - 0 +
+ + - 0 0 0 + - ...

PER vs ORG

LOC vs LOC

ORG vs ORG

P/L vs O/M

P/O vs L/M

P/M vs L/O

Wide-Coverage Spanish Named Entity Extraction
Features

- Sliding Window, codifying ±3 words.

- Primitive Features in the window:
  - Word form
  - PoS (when available)
  - Ortographic Features
    - initial-caps
    - roman-number
    - acronym
    - single-char
    - all-caps
    - contains-dots
    - lonely-initial
    - functional-word
    - all-digits
    - contains-hyphen
    - punctuation-mark
    - 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 (BIO or Open/Close tags)
Weak Rules (1)

From **Open** classifier

```
o0::Caps -- tl1::Out -- w12::NA -- -0.430
   |         |         |
   |         |         -- 1.058
   |         |
   |         -- ol3::allCaps -- -0.859
   |         |
   |         -- -2.203
   |
   -- w0:" -- or1::Caps -- -0.092
      |         |
      |         -- -2.614
      |
      -- ol1::Fun -- -2.799
         |         -- -4.318
```
Weak Rules (2)

From Close classifier

```
or1:Cap -- or3:allCap -- oB:allCap -- 0.473
|     |     |   -- or2:Cap -- -1.366
|     |     |   -- or1:Fun -- or2:Cap -- -0.976
|     |     |   -- 0.252
|     |     |   -- or1:allCap -- -2.191
|     |     |   -- 1.741
```
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: ~65,000 words (4,820 NE)
  - NEC training set: ~700,000 words (61,266 NE)
  - NER training set: ~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

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

<table>
<thead>
<tr>
<th>NEC</th>
<th>#Exs.</th>
<th>#Feat.</th>
<th>#Pos.examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>61,266</td>
<td>22,465</td>
<td>12,976 (21.18%)</td>
</tr>
<tr>
<td>LOCATION</td>
<td>61,266</td>
<td>22,465</td>
<td>14,729 (24.04%)</td>
</tr>
<tr>
<td>ORGANIZ</td>
<td>61,266</td>
<td>22,465</td>
<td>22,947 (34.46%)</td>
</tr>
<tr>
<td>OTHER</td>
<td>61,266</td>
<td>22,465</td>
<td>10,614 (17.32%)</td>
</tr>
</tbody>
</table>
Results on the NER Task (1)

![Graph showing F1 measure vs number of rounds for different methods.]

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACO⁺</td>
<td>89.94%</td>
<td>87.51%</td>
<td>88.71%</td>
</tr>
<tr>
<td>OpenClose</td>
<td>92.42%</td>
<td>91.54%</td>
<td>91.97%</td>
</tr>
<tr>
<td>BIO</td>
<td>92.66%</td>
<td>91.99%</td>
<td>92.33%</td>
</tr>
<tr>
<td>OpenClose⁺I</td>
<td>92.60%</td>
<td>92.14%</td>
<td>92.37%</td>
</tr>
</tbody>
</table>

Wide-Coverage Spanish Named Entity Extraction
Results on the NER Task (2)

<table>
<thead>
<tr>
<th>Subset</th>
<th>#NE</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>length=1</td>
<td>2,807</td>
<td>94.64%</td>
<td>95.65%</td>
<td>95.15%</td>
</tr>
<tr>
<td>length=2</td>
<td>1,005</td>
<td>94.01%</td>
<td>93.73%</td>
<td>93.87%</td>
</tr>
<tr>
<td>length=3</td>
<td>495</td>
<td>91.65%</td>
<td>88.69%</td>
<td>90.14%</td>
</tr>
<tr>
<td>length=4</td>
<td>237</td>
<td>84.81%</td>
<td>84.81%</td>
<td>84.81%</td>
</tr>
<tr>
<td>length=5</td>
<td>89</td>
<td>77.27%</td>
<td>76.40%</td>
<td>76.84%</td>
</tr>
<tr>
<td>length=6</td>
<td>74</td>
<td>81.94%</td>
<td>79.73%</td>
<td>80.82%</td>
</tr>
<tr>
<td>length=7</td>
<td>22</td>
<td>60.00%</td>
<td>54.55%</td>
<td>57.14%</td>
</tr>
<tr>
<td>length=8</td>
<td>22</td>
<td>88.24%</td>
<td>68.18%</td>
<td>76.92%</td>
</tr>
<tr>
<td>length=9</td>
<td>11</td>
<td>80.00%</td>
<td>72.73%</td>
<td>76.19%</td>
</tr>
<tr>
<td>length=10</td>
<td>3</td>
<td>50.00%</td>
<td>33.33%</td>
<td>40.00%</td>
</tr>
<tr>
<td>uppercase</td>
<td>4,637</td>
<td>92.81%</td>
<td>93.25%</td>
<td>93.03%</td>
</tr>
<tr>
<td>lowercase</td>
<td>183</td>
<td>85.40%</td>
<td>63.93%</td>
<td>73.13%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>4,820</td>
<td>92.60%</td>
<td>92.14%</td>
<td>92.37%</td>
</tr>
</tbody>
</table>
### Results on the NEC Task

<table>
<thead>
<tr>
<th>Method</th>
<th>$P$</th>
<th>$R$</th>
<th>$F1$</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most freq</td>
<td>39.78%</td>
<td>39.78%</td>
<td>39.78%</td>
<td>39.78%</td>
</tr>
<tr>
<td>basic</td>
<td>90.19%</td>
<td>84.44%</td>
<td>87.22%</td>
<td>87.51%</td>
</tr>
<tr>
<td>+tw</td>
<td>90.11%</td>
<td>84.77%</td>
<td>87.36%</td>
<td>88.17%</td>
</tr>
<tr>
<td>+gaz</td>
<td>90.25%</td>
<td><strong>85.31%</strong></td>
<td>87.71%</td>
<td>88.60%</td>
</tr>
<tr>
<td>+tw+gaz</td>
<td><strong>90.61%</strong></td>
<td>85.23%</td>
<td><strong>87.84%</strong></td>
<td><strong>88.73%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$P$</th>
<th>$R$</th>
<th>$F1$</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most freq</td>
<td>37.47%</td>
<td>37.47%</td>
<td>37.47%</td>
<td>37.47%</td>
</tr>
<tr>
<td>basic</td>
<td>83.84%</td>
<td>79.39%</td>
<td>81.55%</td>
<td>81.85%</td>
</tr>
<tr>
<td>+tw</td>
<td>85.08%</td>
<td>79.30%</td>
<td>82.09%</td>
<td>82.15%</td>
</tr>
<tr>
<td>+gaz</td>
<td>85.05%</td>
<td>79.57%</td>
<td>82.22%</td>
<td>82.15%</td>
</tr>
<tr>
<td>+tw+gaz</td>
<td><strong>85.32%</strong></td>
<td><strong>79.80%</strong></td>
<td><strong>82.47%</strong></td>
<td><strong>82.31%</strong></td>
</tr>
</tbody>
</table>
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)