Application of Information Extraction in Large Semi-Structured Text

Valmir Macário (Cin\UFPE)
Ricardo Prudêncio (Cin\UFPE)
Francisco Carvalho (Cin\UFPE)
Leandro Torres (Capital Login)
Laerte Rodrigues (Capital Login)

1. Introduction
2. Information Extraction
3. Information Extraction on Official Journal
4. Experiments
5. Results
6. Conclusions
7. Future Works

Introduction

- A great amount of valuable information is stored in digital repositories

- Legible by humans but hardly manipulated by computer machines.

Introduction

- Important automatically extract information on these repositories in order to support specific uses
- Information Extraction (IE) systems are able to extract specific information of interest

Information Extraction

- The main objective of an IE system is to recognize pieces of information from texts which correspond to data fields required by the users.

- IE systems building depends on the kind of the texts being tackled

Information Extraction

- Document Types:
  - Free texts:
    - Unrestricted Natural language
    - Any formatation or regular pattern
    - Natural Language Processing techniques

  - Structured texts:
    - Suitable for machine computers
    - Rigid Format
    - Uniform rules

Information Extraction

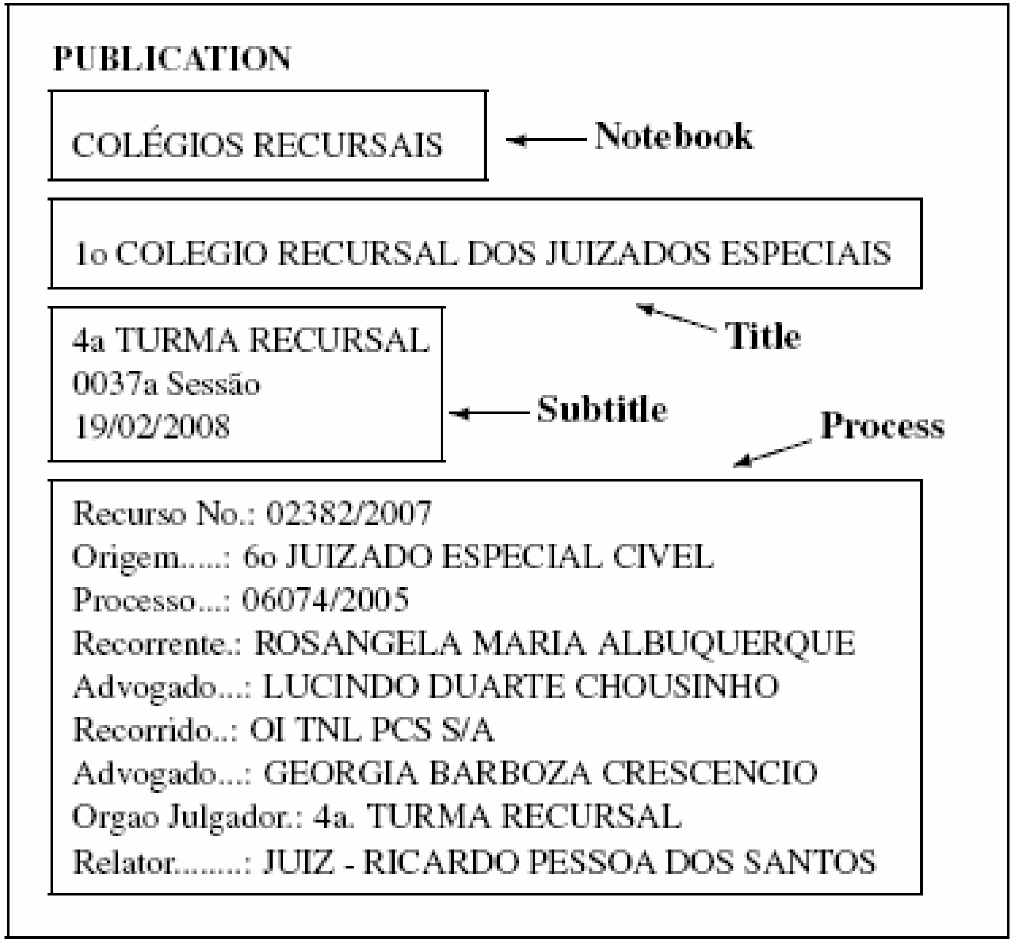
- Document Types:
  - Semi-structured texts:
    - Some structure
    - Missing fields
    - Replaced order of fields
    - Lack of delimiters between fields
    - Abbreviate words
    - Machine Learning (ML) or Natural Language Processing techniques

Information Extraction on Official Journal

- The current work presents an IE system that extracts publications available in official journals

- Machine Learning Approach

Information Extraction on Official Journal



PUBLICATION

COLÉGIOS RECURSAIS ← Notebook

1o COLEGIO RECURSAL DOS JUIZADOS ESPECIAIS

Title

4a TURMA RECURSAL
0037a Sessão
19/02/2008

← Subtitle

Process

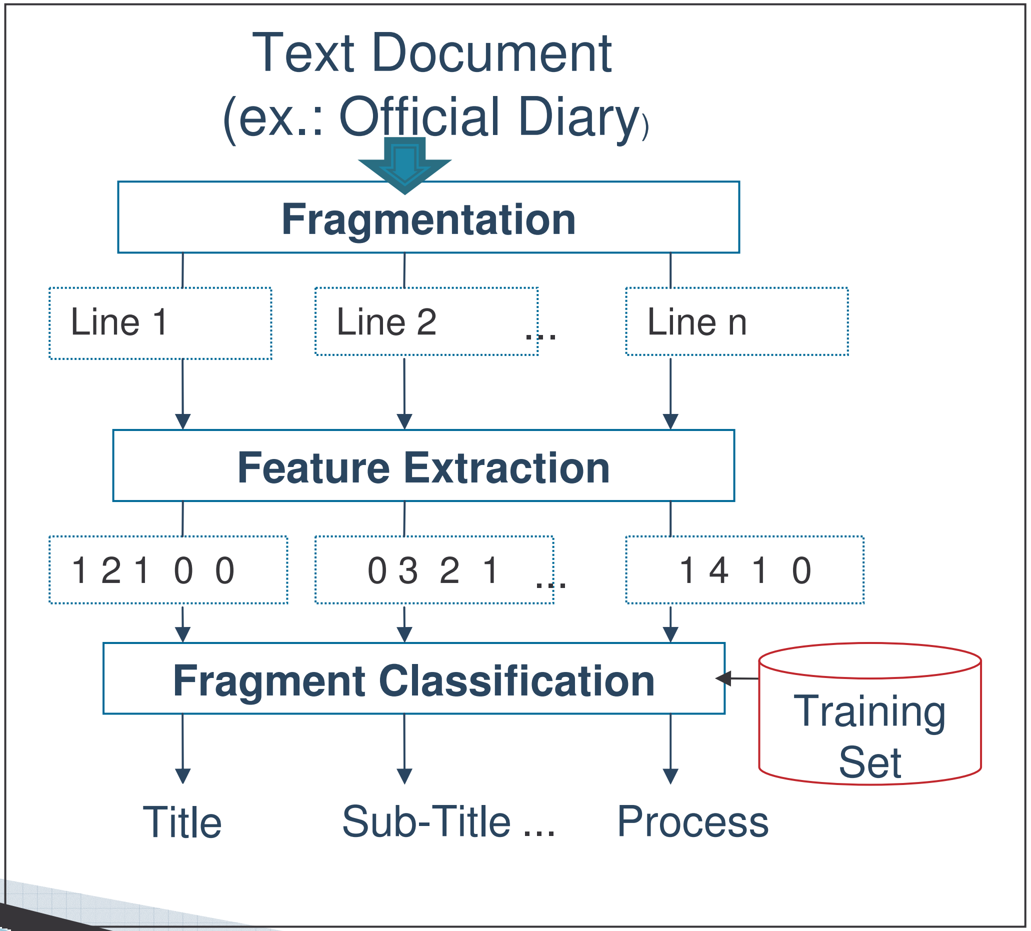
Recurso No.: 02382/2007
Origem.....: 6o JUIZADO ESPECIAL CIVEL
Processo....: 06074/2005
Recorrente.: ROSANGELA MARIA ALBUQUERQUE
Advogado...: LUCINDO DUARTE CHOUSINHO
Recorrido..: OI TNL PCS S/A
Advogado...: GEORGIA BARBOZA CRESCENCIO
Orgao Julgador.: 4a. TURMA RECURSAL
Relator........: JUIZ - RICARDO PESSOA DOS SANTOS

Information Extraction on Official Journal

- Difficulties to extract information from official journals:
  - Fields may present very similar patterns
    - “Edital de Intimação” can be a process or a subtitle
  - Absent fields
  - Presence of abbreviated patterns
    - Process, Proc., Proc
  - Ortographic errors

Information Extraction on Official Journal

- Generic architecture of the IE system:



Information Extraction on Official Journal

- Fragmentation:
  - Divide the text in little pieces
  - Text delimiters: marks, white spaces, end-of-line, characters, paragraph characters, among others.

- End-of-Line

Information Extraction on Official Journal

- Feature Extraction:
  - *Vocabulary:* 180 words defined by a domain expert
  - *Regular Expression:* patterns presented in text that can be represented as regular expressions (280 regular expressions).
  - *General:* 15 features based on the Bouckaert's work (Eg. Starts with uppercase, contains numbers or enumerations)

Fragment Classification

- *Possible classes (10 classes):*
  - *iprocess,process, ititle, title, isubtitle, subtitle, notebook, city, nil and blank*

- Sliding Window (SW) approach with overlapping

- Three classifiers:
  - PART
  - Naive Bayes
  - Support Vector Machine

Experiments Description

- Official Journal published by the State of Pernambuco, Brazil

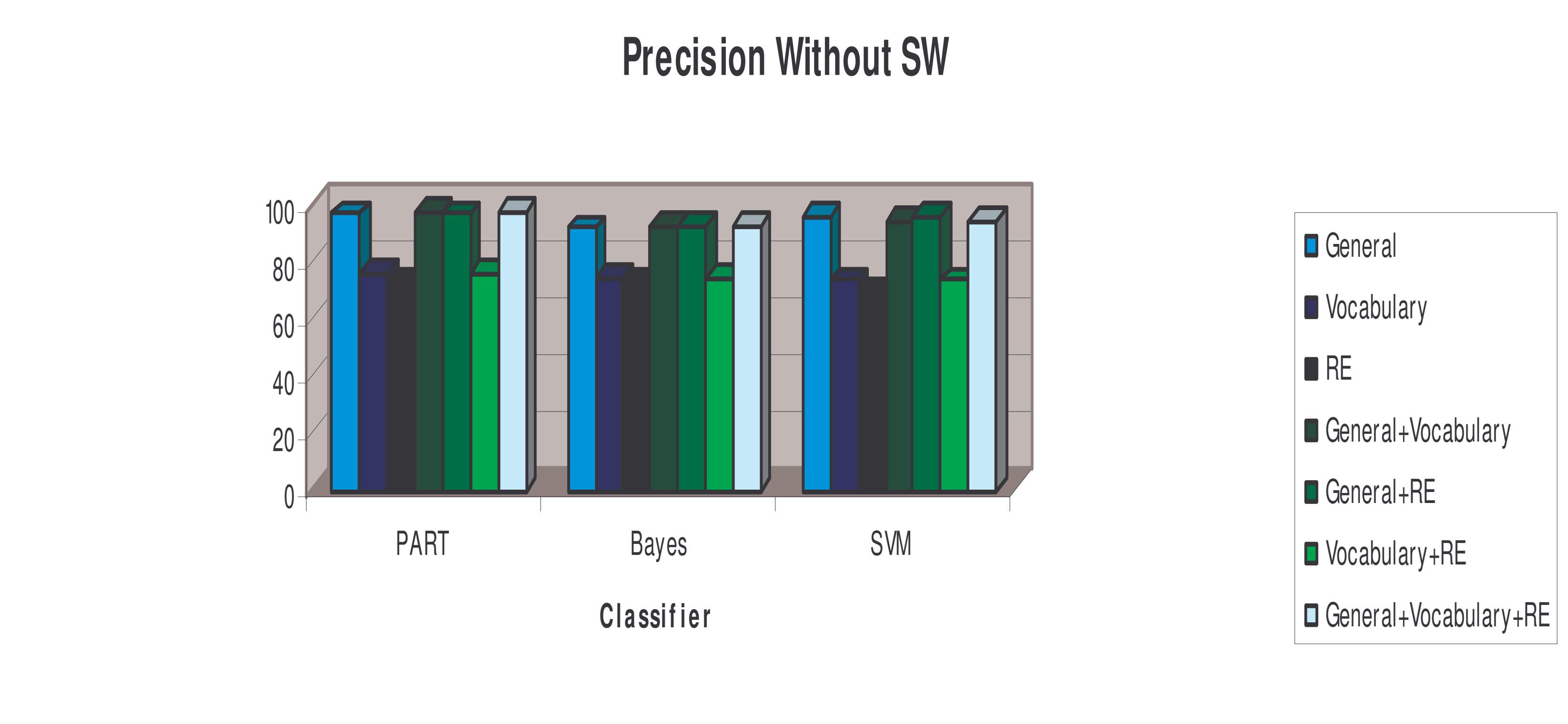
- Pages published from 8 to 14 February, 2008. 22,770 Lines.

- IE system was evaluated for 21 different scenarios (i.e., different combinations of feature sets *versus classifiers).*

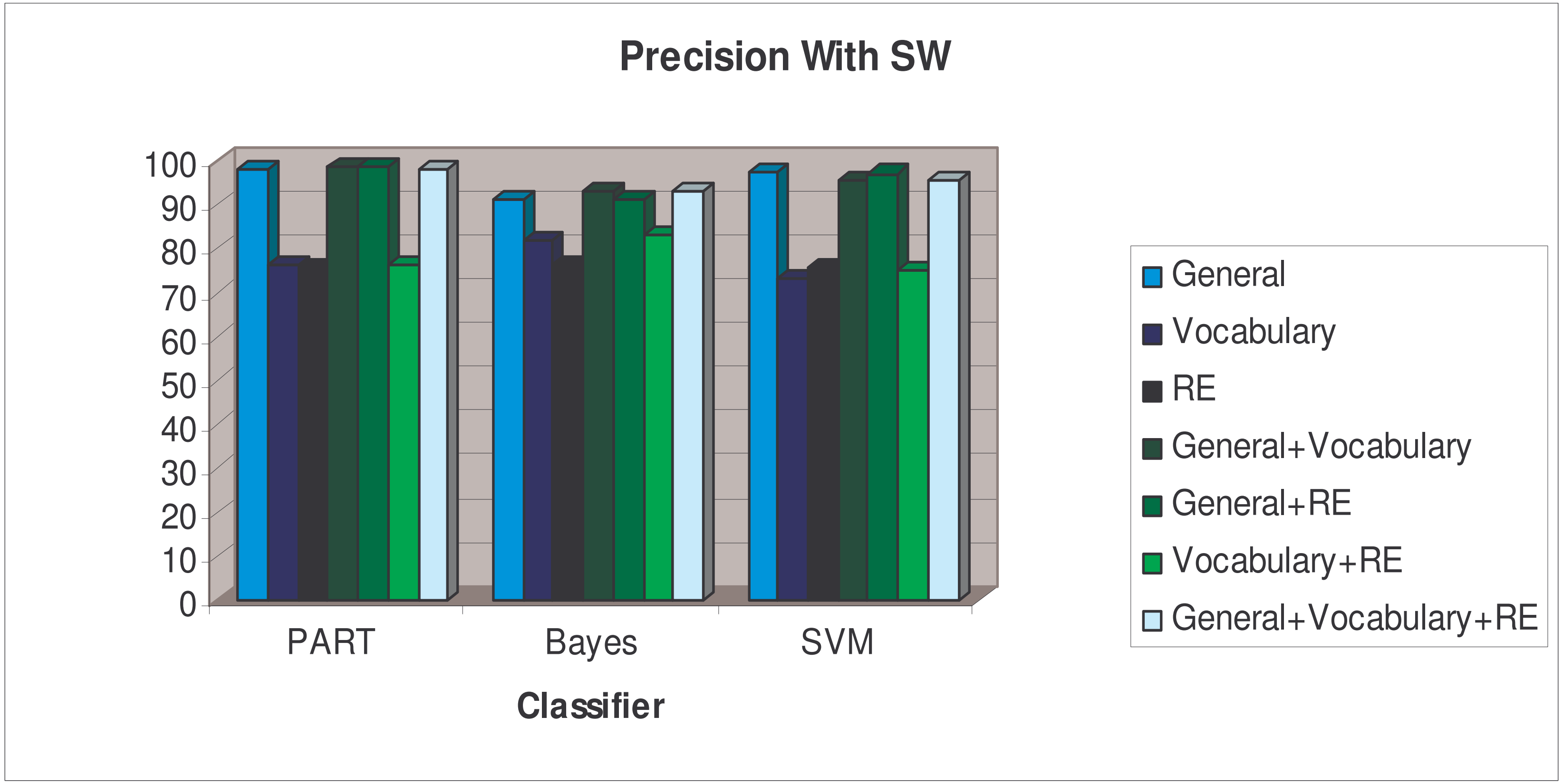
Experiments Description

- (1) General (15 features)
- (2) Vocabulary (180 features)
- (3) Regular Expression (RE) (280 features)
- (4) General + Vocabulary (195 features)
- (5) General + RE (295 features)
- (6) Vocabulary + RE (460 features)
- (7)General + Vocabulary + RE (475 attributes).

Results

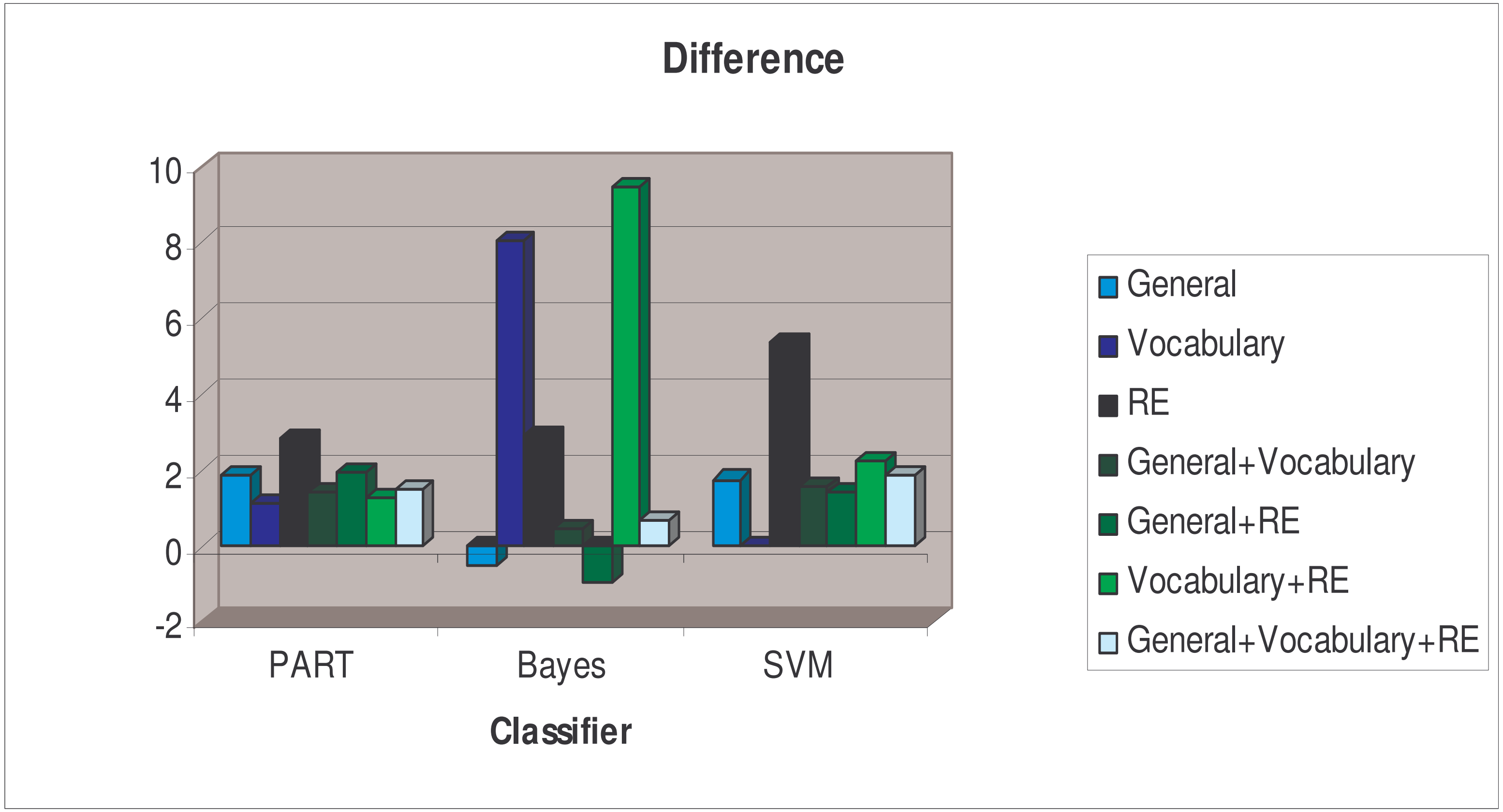

Precision Without SW
100
80
60
40
20
0
PART
Bayes
SVM
Classifier
General
Vocabulary
RE
General+Vocabulary
General+RE
Vocabulary+RE
General+Vocabulary+RE

Results



Precision With SW
100
90
80
70
60
50
40
30
20
10
0
PART
Bayes
SVM
Classifier
General
Vocabulary
RE
General+Vocabulary
General+RE
Vocabulary+RE
General+Vocabulary+RE

Results


Difference
10
8
6
4
2
0
-2
PART
Bayes
SVM
Classifier
General
Vocabulary
RE
General+Vocabulary
General+RE
Vocabulary+RE
General+Vocabulary+RE

Results

Table 2. Average precision for each classifier

Classifier	Average Precision without SW	Average Precision with SW
PART	87.25	**88.97**
Bayes	84.14	87.02
SVM	85.03	87.08

Table 3. Average precision for each feature set.

Feature Set	Average Precision without SW	Average Precision with SW
General	94.47	95.50
Vocabulary	73.98	77.07
RE	72.11	75.85
General+Vocabulary	94.53	94.27
General+RE	94.77	95.58
Vocabulary +RE	73.98	78.30
General+Vocabulary + RE	94.49	**95.85**

Conclusion

- The IE system developed reveal to be adequate for our purpose
- The feature set influences the System perfomance
- Sequential information improves the classification precision

Future Work

- Automatic feature selection to construct the feature sets.
- Evaluate sequential learning algorithms, such Hidden Markov Models and Conditional Random Fields.

Bibliography

- Appelt, D. and Israel, D. Introduction to Information Extraction Technology. *IJCAI–99 Tutorial*, Stockholm, Sweden, 1999.
- Feldman, R., Rosenfeld, B. and Fresko, M. TEG: a hybrid approach to information extraction. *Knowledge and Information Systems* 9(1):1–18, 2006.
- Laferty, J., McCallum, A. and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*. 2001.
- Bouckaert, R. R. Low level information extraction: a bayesian network based approach. In *TextML*, 2002.
- Silva, E. F. A., Barros, F. A. and Prudêncio, R. B. C. A Hybrid Machine Learning Approach for Information Extraction. In *Proceedings of the 6th International Conference on Hybrid Intelligent Systems*, Auckland, New Zealand, 2006.