Adaptive Information Extraction from Web Pages by Supervised Wrapper Induction

by

Rinaldo Lima (rjl4@cin.ufpe.br),
Phd. Fred Freitas (fred@cin.ufpe.br) and
Phd. Bernard Espinasse (espinasse@lsis.org)

Recife, Novembre 2009

Centro de Informática U·F·P·E

Friends meeting at the LSIS Labs.

Universidade Federal de Pernambuco

Outline

1. What is Information Extraction?
2. Basic concepts in IE: *type of texts* and *extractions*
3. Wrapper Induction: definition and examples
4. Boosted Wrapper Induction
5. IE as a Classification Problem
6. Proposal of an IE architecture
7. Experimental Results
8. Conclusions and Future Work

What is Information Extraction?

Information Extraction (EI) is the task of identifying the relevant information fragments of text from larger documents
appropriate formats for future use
[auchak, 2004].

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

IE

Extraction Template

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	FreeSoft.

Type of Texts in IE

Centro de Informática U·F·P·E

Text paragraphs without formatting

(a)

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artif
Mell
indu is
M.S.
computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Free

Grammatical sentences and some formatting & links

(b)

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was
the f
Intell
Mint
proje
Instit
Carn
Princ
taugh

• Press
Contact

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

Semi-structured

(c) Non-grammatical snippets, rich formatting & links

Barto, Andrew G. (413) 545-2109 barto@cs.umass.edu CS276
Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning
cont

Berger, ass.edu CS344
Assi

Brock, C ass.edu CS246
Assi

Clarke, ass.edu CS304
Prof
Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

Semi-structured

(d) Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	**Logic Programming**	**Natural Language Generation**	**Complexity Analysis**	**Neural Networks**	**Games**
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz and Gerhard Lakemeyer*	131: A Comparative Study of Logic Programs with Preference *Torsten Schaub and Kewen*	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation *Adnan Darwiche and Pierre Marquis*	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing

Structured

What to Extract: Single/Multiple slots

Jack Welch *will retire as* ***CEO*** *of General Electric tomorrow. The top role at the Connecticut company will be filled by* ***Jeffrey Immelt****.*

(a) Single entity (**Template filling**)

(b) Binary relationship (**Relation Extraction**)

(c) N-ary record (**Scenario Extracti**

Person: Jack Welch

Person: Jeffrey Immelt

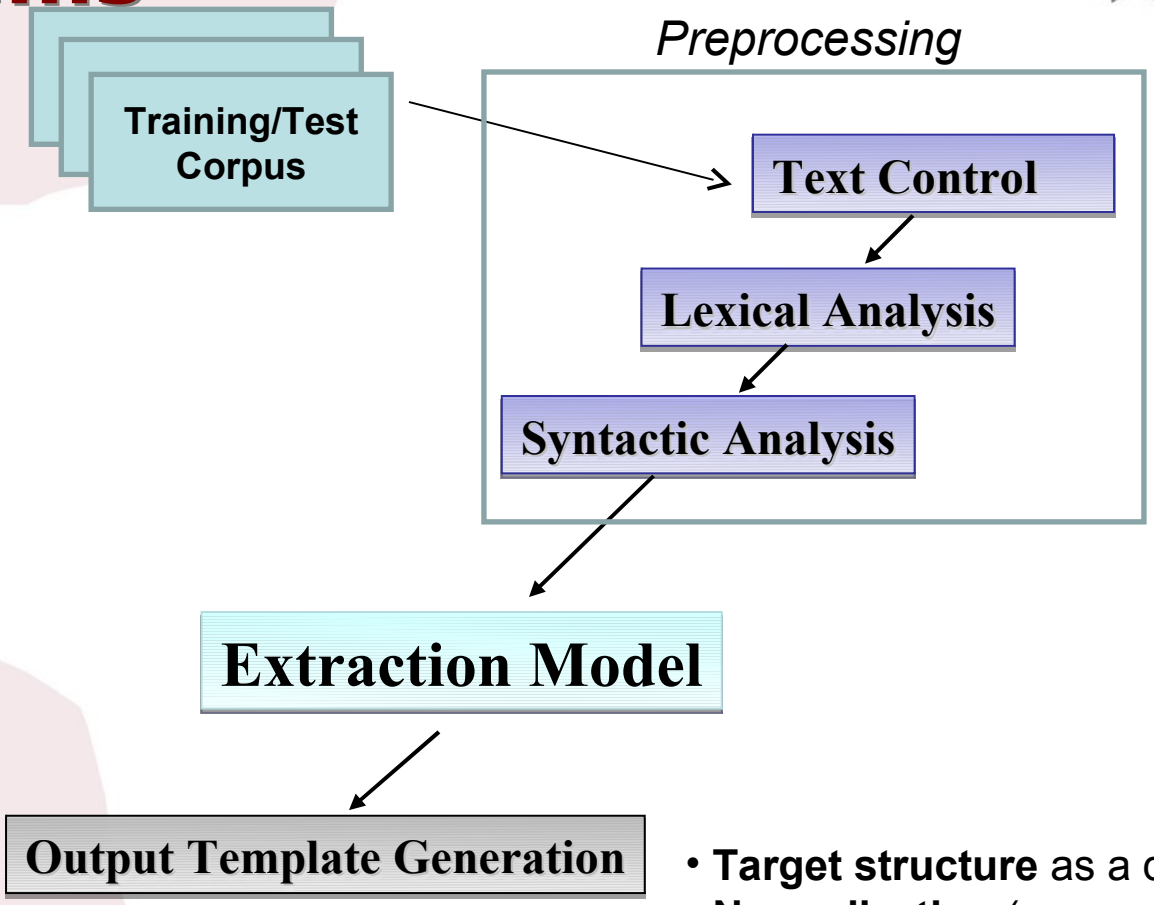
Location: Connecticut

Relation: Person-Title
Person: Jack Welch
Title: CEO

Relation: Company-Location
Company: General Electric
Location: Connecticut

Relation: Succession
Company: General Electric
Title: CEO
Out: Jack Welsh
In: Jeffrey Immelt

Typical Architecture of IE Systems



Tokenization and/or Feature Extraction

POS Tagging

Sentece Spliting

- **Target structure** as a database relation
- **Normalization** (same representation for date and numbers, for example)
- **Instance unification** (co-reference resolution)

Boosted Wrapper Induction

Boosted Wrapper Induction (BWI) [Freitag & Kushmerick, 2000]

- A document is treated as a sequence of **tokens**, and the IE task is to identify the **boundaries** of each type of information to be extracted

- It learns extraction rules composed only of **simple contextual patterns:**
  - **common prefixes and suffixes** of the text occurring immediately before (or after) the text fragments to be extracted

Boosted Wrapper Induction
[Freitag & Kushmerik, 2000]

Examples:

(1) The folowing prefix and sufix

- < **[<href=“] , [http]** > determine a **fore-dectector** of an URL
- < **[.html], [“>]** > determine an **after-detector** of an URL

Boundaries

http://xyz.com/index.html de **<a** **href=** http**://xyz.com/index.html >**

(2)

top-scoring boundary detectors

$F_1 = \langle$[time :], [<Num>]$\rangle$

$A_1 = \langle$[], [- <Num> : <*> <Alph>

example

... Time: 2:00 - 3:30 PM

Fig. fore and after detectors **<F, A>** generated by the BWI algorithm [Freitag & Kushmerick, 2000].

IE as a Classification Problem: Boundary Detectors

- The IE task is to identify the **boundaries** that indicate the beginning and the end of each field.
- Formally, **boundary detector d = <p, s>** is a pair of **prefix p** and **suffix s** patterns that matches a boundary **i** if p matches the tokens before ***i*** and s matches the tokens after ***i***.

- Associated with every detector **d** is a numeric **confidence value C_d**.

- A **wrapper W = < F, A, H>** consists of two sets **F** and **A** of detectors and **H(k)** reflects the probability that a field has length ***k***.

- To perform **extraction using W**, every boundary ***i*** in a document is first given a **fore** and a **aft score**. W then classifies **text fragment <i, j>** as follows:

$$W(i, j) = \begin{cases} 1, & \text{if } F(i)\ A(j)\ H(j - i) \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

IE as a Classification Problem: *Boosting*

BWI uses **boosting** to generate and **combine** the predictions from numerous extraction patterns.

procedure BWI(example sets S and E)
  $F \leftarrow$ AdaBoost(LearnDetector, S)
  $A \leftarrow$ AdaBoost(LearnDetector, E)
  $H \leftarrow$ field length histogram from S and E
  return wrapper $W = \langle F, A, H \rangle$

→ **Boosting** is a procedure to improving the performance of a "**weak**" machine learning algorithm by repeatedly applying it to the training set, at each iteration modifying training example **weight**s to emphasize examples on which the weak learner has done poorly in previous iterations [Schapire & Singer,1998].

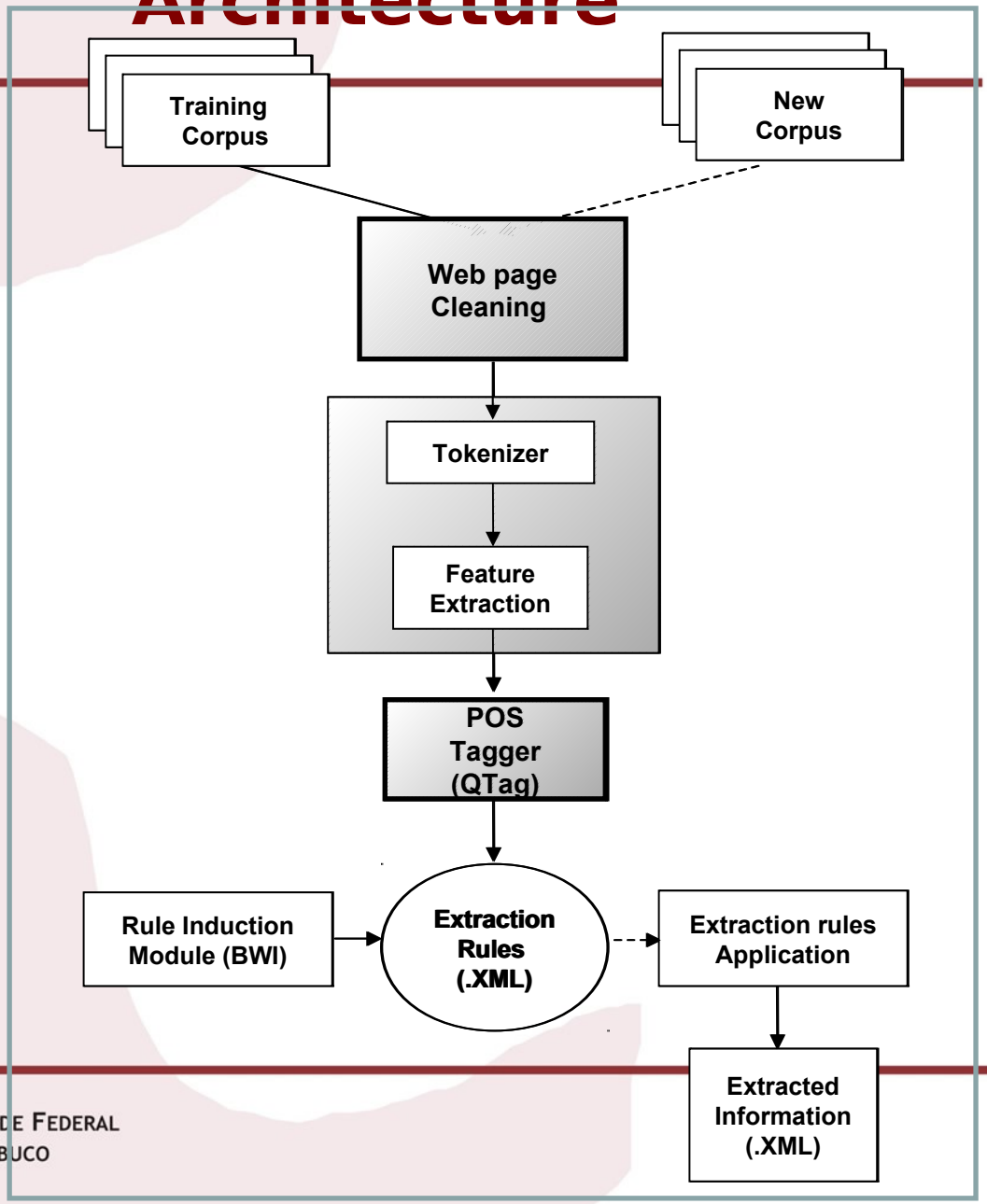
Proposal of an IE Architecture: WEPAIES

- WEPAIES integrates the IE system **TIES** (*Trainable Information Extraction System*) , developed at ITC-irst (Instituto Trentino di Cultura).

- TIES is a Java implementation of the BWI algorithm. The extractors are based on the Boosted Wrapper Induction (BWI) algorithm.

- TIES automatically learns rules from a previously annotated corpus with a predefined set of tags (template).

WEPAIES Architecture



Tokenization, Feature Extraction and ~~POS tagging: Preprocessing~~

Input text: "CALL FOR PAPERS (v4)"

line

div/token

id	Text	type	pos	start	len	alpha	upper_	symb_	single_	lower_	num_t
22	CALL	word	NN	62	4	true	true				
24	FOR	word	IN	67	3	true	true				
26	PAPERS	word	NNS	71	6	true	true				
28	(	sym	(	78	1			true	true		
29	v	word	NN	79	1	true			true	true	
30	4	num	CD	80	1				true		true
31	)	sym	)	81	1			true	true		

1	2	3	4	5	6	7	8	9	10

Tokenization

Tokenization, Feature Extraction and POS tagging: Preprocessing

Input text: "CALL FOR PAPERS (v4)"

line

div/token

id	Text	type	pos	start	len	alpha	upper_	symb_	single_	lower_	num_t
22	CALL	word	NN	62	4	true	true				
24	FOR	word	IN	67	3	true	true				
26	PAPERS	word	NNS	71	6	true	true				
28	(	sym	(	78	1			true	true		
29	v	word	NN	79	1	true			true	true	
30	4	num	CD	80	1				true		true
31	)	sym	)	81	1			true	true		

1	2	3	4	5	6	7	8	9	10

Default Feature Extraction

Tokenization, Feature Extraction and POS tagging: Preprocessing

Input text: "CALL FOR PAPERS (v4)"

line

div/token

id	Text	type	pos	start	len	alpha	upper_	symb_	single_	lower_	num_t
22	CALL	word	NN	62	4	true	true				
24	FOR	word	IN	67	3	true	true				
26	PAPERS	word	NNS	71	6	true	true				
28	(	sym	(	78	1			true	true		
29	v	word	NN	79	1	true			true	true	
30	4	num	CD	80	1				true		true
31	)	sym	)	81	1			true	true		

1	2	3	4	5	6	7	8	9	10

Extended Feature Extraction

Tokenization, Feature Extraction e ~~POS tagging: Pré-processamento~~

Input text: "CALL FOR PAPERS (v4)"

line

div/token

id	Text	type	pos	start	len	alpha	upper_	symb_	single_	lower_	num_t
22	CALL	word	NN	62	4	true	true				
24	FOR	word	IN	67	3	true	true				
26	PAPERS	word	NNS	71	6	true	true				
28	(	sym	(	78	1			true	true		
29	v	word	NN	79	1	true			true	true	
30	4	num	CD	80	1				true		true
31	)	sym	)	81	1			true	true		

1	2	3	4	5	6	7	8	9	10

POS Tagging

Tag	Mean
NN	Singular Noun
IN	Preposition
NNS	Plural Noun
CD	Cardinal number

Testbed Corpus: Seminars

Frequency distribution of the Seminar corpus

Seminars (485 docs)	*Location*	*Speaker*	*Stime*	*Etime*	*Non-Entity*
	643	754	980	433	157.647

<doc id='276' filename='cmu.cs.proj.vision-273_0'><0.25.4.84.12.33.15.???@???.0>
Type: cmu.cs.proj.vision
Topic: Sanderson group seminar
Dates: 27-Apr-84
Time: **<stime>**2:30**</stime>**
PostedBy: ??? on 25-Apr-84 at 12:33 from ???
Abstract:
<speaker>Alberto Elfes**</speaker>** will be speaking about "A Wide-Beam Sonar Mapping System" on Friday the 27th in **<location>**WeH 4623**</location>** at **<stime>**2:30**</stime>.**
</doc>

Annotated document from the Seminar Corpus

Testbed Corpora: Jobs and CFP

Jobs Announcements Corpus (300 docs)

JOBS	Platform	Language	Area	City	State	Application
	709	851	1005	659	452	590
	Title	Recruiter	Post date	Country	Salary	Req-years-e
	457	312	302	345	141	166
	Company	Des-years_e	Req-degree	Des-degree	Id	
	298	43	83	21	304	

Call for Papers Corpus Pascal Challenge 2005 (400 docs)

Annotation Type	Corpus Frequency			
	Train	%	Test	%
workname	543	11.8	245	10.8
workacro	566	12.3	243	10.7
workhome	367	8.0	215	9.5
workloca	457	10.0	224	9.9
workdate	586	12.8	326	14.3
workpape	590	12.9	316	13.9
worknoti	391	8.5	190	8.4
workcame	355	7.7	163	7.2
confname	204	4.5	90	4.0
confacro	420	9.2	187	8.2
confhome	104	2.3	75	3.3
Total	**4583**	**100**	**2274**	**100**

Pascal Challenge on Evaluating Machine Learning for IE (2005)

Goal: Provide a *testbed* for comparative evaluation of ML-based IE. (Ireson, 2005)

Standardisation

- Data
  - Partitioning
  - Same set of features
    - Corpus pre-processed using Gate
    - No features allowed other than the ones provided
- Explicit Tasks
- Evaluation Metrics

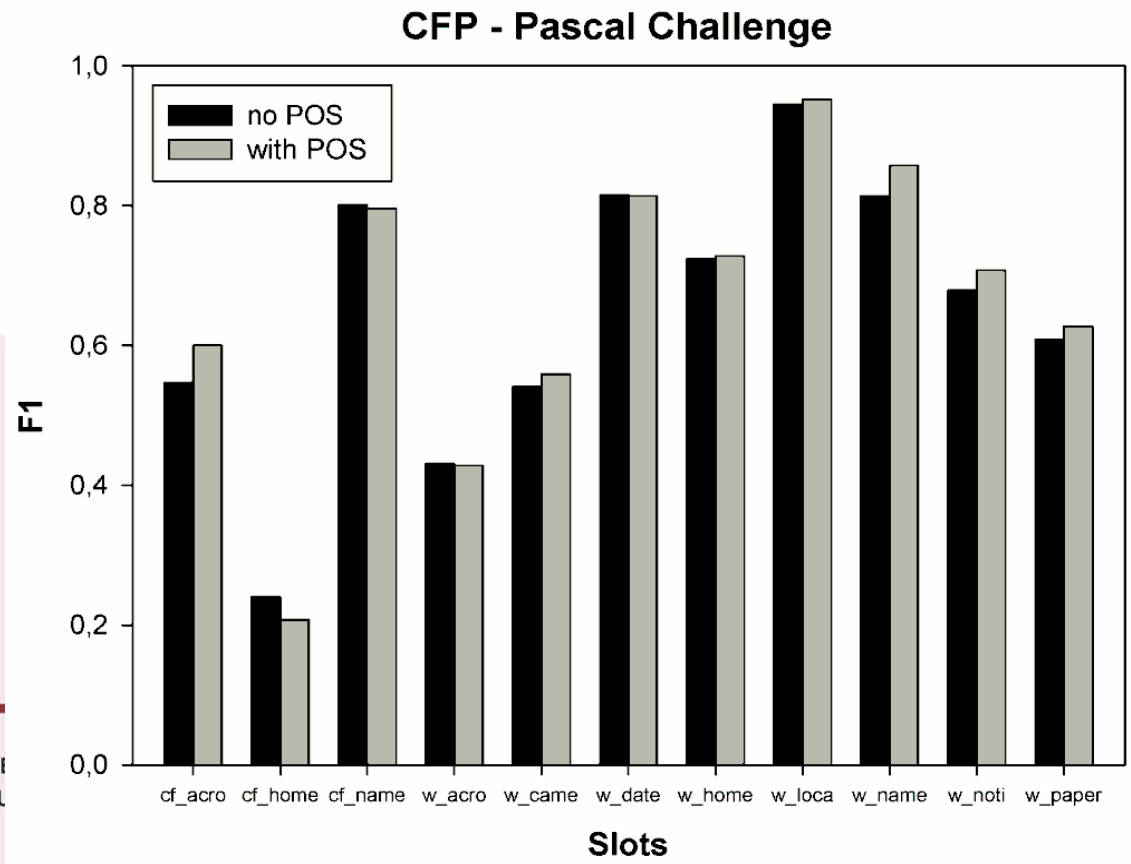
WEPAIES: POS Influence on Corpora

No POS

Corpus	Prec	Recall	F1
Seminars	0,974	0,953	0,963
Jobs	0,945	0,778	0,853
CFP	0,891	0,571	**0,696**

With POS

Corpus	Prec	Recall	F1
Seminars	0,971	0,964	0,967
Jobs	0,939	0,780	0,853
CFP	0,896	0,591	**0,712**



Comparative Evaluation

System	Description
(LP)2	It uses **Shallow NLP** techniques to generalize rules beyond the flat word structure by a **covering** algorithm. The champion of the Pascal Challenge, 2005.
Rapier	Single-slot IE system for free texts that uses POS Information and **wordnet synsets**. It is based on Inductive Logic Programming [Califf & Mooney, 1999].
GATE-SVM	Performing supervised token classification based on a variant of the SVM with **uneven margins** [Li et al., 2003]
Yaoyong	Predecessor of GATE-SVM.
SIE	Performing supervised Token classification using a **filtering instance** technique in its preprocessing phase [Giuliano, 2004]

Comparative Evaluation: SEMINARS Corpus

	speaker	*location*	*stime*	*etime*	***All Slots***
WEPAIES	**86,2**	**88,8**	93,9	**96,7**	**91,4**
SIE	-	-	-	-	86,6
GATE-SVM	69,0	81,3	94,8	92,7	86,2
$(LP)^2$	77,6	75,0	**99,0**	95,5	86,0
Rapier	53,0	72,7	93,4	96,2	77,3

Comparative Evaluation: JOBS Corpus

Slot	(LP)^2	GATE_SVM	**WEPAIES**	Rapier
id	**100,0**	97,7	98,1	97,5
title	43,9	49,6	**67,4**	40,5
company	71,9	77,2	**78,9**	69,5
salary	62,8	86,5	**89,2**	67,4
recruiter	80,6	78,4	**86,1**	68,4
state	86,7	92,8	**96,9**	90,2
city	93,0	95,5	**96,5**	90,4
country	81,0	96,2	**98,8**	93,2
language	**91,0**	86,9	88,5	80,6
plataform	80,5	80,1	**86,9**	72,5
application	**78,4**	70,2	73,1	69,3
area	**66,9**	46,8	**51,6**	42,4
req_y_exp	68,8	80,8	**86,4**	67,1
des_y_exp	60,4	81,9	**89,9**	87,5
req_degree	84,7	**87,5**	78,6	81,5
des_degree	**65,1**	59,2	**47,6**	72,2
post date	99,5	99,2	**100,0**	99,5
All slots	**84,1**	**80,8**	**83,8**	**75,1**

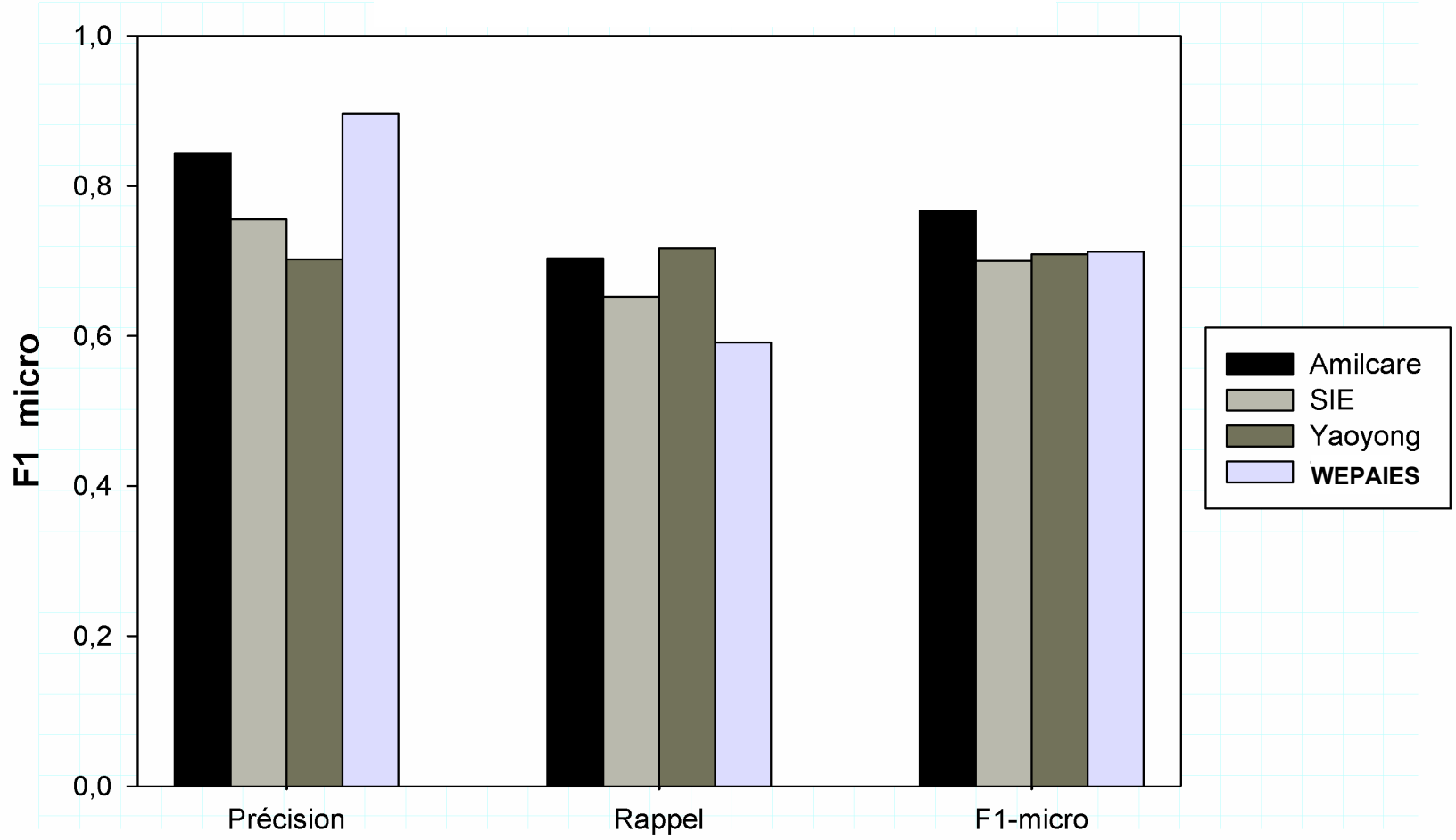
Comparative Evaluation: CFP

Comparative Slots Performance on the CFP

P – Precision, R – Recall, F – F-Measure.

		WORKSHOP								CONFERENCE		
System	**Sc.**	**name**	**acro**	**date**	**home**	**loca**	**pape**	**noti**	**came**	**name**	**acro**	**home**
Amilcare (LP)[2]	P	0,656	0,887	0,769	0,864	0,621	0,876	0,889	0,876	0,792	0,922	0,656
	R	0,241	0,884	0,632	0,619	0,402	0,851	0,889	0,865	0,422	0,888	0,280
	F	0,352	**0,865**	0,694	0,721	0,488	**0,864**	**0,889**	**0,870**	0,551	**0,905**	**0,393**
Yaoyong	P	0,629	0,738	0,810	0,656	0,611	0,719	0,867	0,764	0,649	0,619	0,368
	R	0,539	0,523	0,666	0,870	0,674	0,763	0,821	0,736	0,411	0,348	0,093
	F	0,580	0,612	0,731	**0,748**	0,641	0,740	0,843	0,750	0,503	0,445	0,149
SIE	P	0,852	0,733	0,850	0,672	0,812	0,841	0,921	0,911	0,795	0,667	0,556
	R	0,539	0,259	0,451	0,419	0,406	0,617	0,795	0,687	0,344	0,235	0,067
	F	0,660	0,383	0,589	0,516	0,542	0,712	0,853	0,783	0,481	0,348	0,119
WEPAIES	P	0,889	0,906	0,918	0,718	0,990	0,906	0,925	0,849	0,953	0,930	0,706
	R	0,825	0,275	0,729	0,735	0,916	0,477	0,569	0,414	0,691	0,443	0,122
	F	**0,856**	0,422	**0,813**	0,726	**0,952**	0,625	0,705	0,556	**0,801**	0,600	0,209

Comparative Evaluation : CFP



Precision, Recall and F-Measure comparison on the CFP Corpus

Conclusion and Future Work

Conclusion:

- The results obtained allow us to conclude that the newly extended TIES system, an adaptive IE system based on supervised wrapper induction is comparable with other state-of-the-art IE systems on traditional IE tasks.
- POS Information is more helpful when our architecture is applied on non structured (free) texts.

Future work :

- the improvement of WEPAIES architecture by including new supervised machine learning algorithms, such as **Support Vector Machines** and **C4.5** as learning components for new IE wrappers.
- is related to the *tokenizer and feature extraction* modules in which we intend to perform the following NLP subtasks, i.e., **NER** and **Chunking Analysis (for English)** and **POS tagging for Portuguese** and **French languages**.

References

[Califf, 1999] CALIFF, M., MOONEY, R. *Relational Learning of Pattern-Match Rules for Information Extraction*, in Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99).

[Espinasse, 2007] ESPINASSE, B., FOURNIER, S. & FREITAS, F. *Agent and Ontology based Information Gathering on Restricted Web Domains with AGATHE.* Domaine Universitaire de St Jerôme. Marseille, France. 2007.

[Freitag, 2000] FREITAG, D. & KUSHMERICK, N. *Boosted wrapper induction.* In Proceedings of 17th National Conference on Artificial Intelligence. pp.577-583. (2000).

[Kauchak, 2004] Kauchak, D., SMARR, J. & ELKAN, C. *Sources of success for boosted wrapper induction.* The Journal of Machine Learning Research, Vol.5, pp.499-527. MA: MIT Press. (2004).

[Ireson, 2005] IRESON, N. et al. Evaluating machine learning for information extraction. In Proceedings International Conference on Machine Learning. (2005)

[LI ET, 2004] LI Y., BONTCHEVA K., CUNNINGHAM H.: SVM BASED LEARNING SYSTEM FOR INFORMATION EXTRACTION. DETERMINISTIC AND STATISTICAL METHODS IN MACHINE LEARNING 2004: 319-339, 2004.

Publications from this work

(1) 25th Symposium On Applied Computing – SAC 2010 – Switzerland

An Adaptive Information Extraction System based on Wrapper Induction with POS Tagging

Rinaldo Lima
Centro de Informática, UFPE
50740-540 Cidade Universitária,
Recife, PE, Brazil
rjl4@gmail.cin.ufpe.br

Bernard Espinasse
LSIS UMR CNRS 6168
Domaine Universitaire de St Jerôme,
F-13997, Marseille Cedex 20, France
bernard.espinasse@lsis.org

Fred Freitas
Centro de Informática, UFPE
50740-540 Cidade Universitária,
Recife, PE, Brazil
fred@cin.ufpe.br

(2) IADIS - Internation Conference - WWW/Internet 2009 - Rome - Italy

WEPAIES: A Web Pages Adaptive Information Extraction System Based on Wrapper Induction with POS Tagging

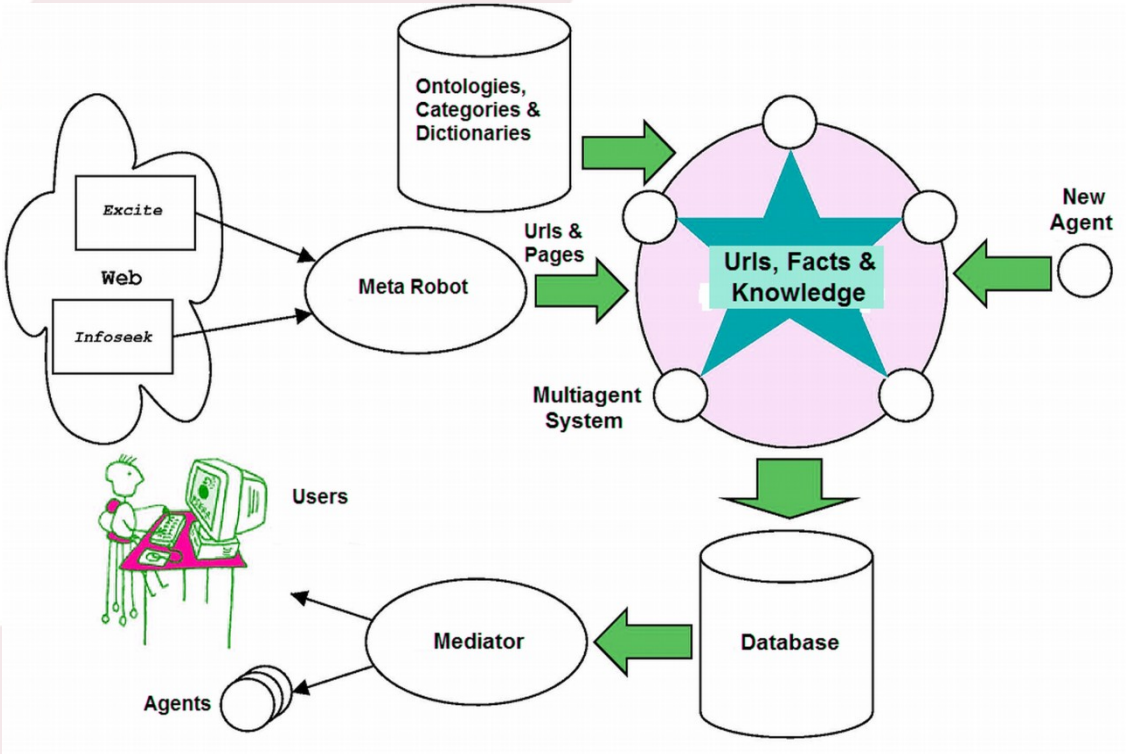
(3) My Master Dissertation

Extraction d'Information Adaptative de Pages Web par Induction Supervisée d'Extracteurs

Adaptive Information Extraction from Web Pages by Supervised Wrapper Induction

Questions?

Context : MasterWeb/AGATHE Systems



It is a generic software architecture for the development of information gathering systems on the Web, for restricted domains, based on software agents that cooperate and exploit ontologies associated to these restricted domains.

In [Espinasse & Freitas, 2007], the following improvements were suggested:

→ Integration of **Machine Learning and Natural Language Processing (NLP)** techniques to accelerate the tasks of extraction and classification to improve information gathering tasks.