Extraction automatique de relations par ontologies et programmation logique inductive

Bernard Espinasse[1] - Rinaldo Lima[2]

[1] LSIS UMR CNRS, Aix Marseille Université, Marseille (AMU), France
bernard.espinasse@lsis.org

[2] Federal Rural University of Pernambuco (UFRPE), Recife, Brazil
rinaldo.jose@ufrpe.br

Janvier 2017

Sommaire

1. Introduction et Motivation

L'extraction d'information (EI)

Extraction Information (EI) composée de 2 tâches principales:

-La reconnaissance d'entités nommées (REN): extraire des instances d'entités nommées, Ex. des noms de personnes, de lieux;

-L'extraction de relations (RE): extraire des relations entre ces entités nommées

Soit la phrase:

"American saxophonist David Murray recruited Amidu Berry"

- Extraction des entités nommées :

"David Murray" et **"Amidu Berry"**

- Extraction des relations:

CITIZEN(David Murray, American**)** et

HIRE(David Murray, Amidu Berry**)**

EI et ressources sémantiques: OBIE

- Pour être plus **précis**, les systèmes d'IE doivent exploiter plus de **ressources sémantiques** (Nédellec & Nazarenko, 2005).
- Emergence de l'EI basée sur des **ontologies** - ***Ontology-Based Information Extraction – OBIE*** (Wimalasuriya et Dou, 2010) :
  - *Ontologie en entrée* : processus d'extraction guidé par une ontologie avec une annotation sémantique des textes à traiter
  - *Ontologie en sortie*: utilisation d'une ontologie pour représenter et stocker les informations extraites par peuplement d'une ontologie
  - L'OBIE permet aussi :
  - D'exploiter un traitement du langage naturel en profondeur
  - De générer automatiquement des contenus sémantiques pour le Web sémantique (Wu and Weld, 2008)

EI et apprentissage automatique

- Pour être plus **rapidement développés** et **adaptables** à d'autres domaines d'application, les systèmes d'EI utilisent des **techniques d'apprentissage automatique**.
- ***L'apprentissage supervisé statistique*** largement utilisé :
  - **REN**: très bonne performance, autour de 90%,
  - **ER**: performance très nettement inférieure (Giuliano et al., 2007) (Bach et Badaskar, 2007), et peu de progrès réalisé depuis un certain temps.
- Pour la RE, une alternative à l'apprentissage supervisée *statistique* est ***l'apprentissage supervisé symbolique***, avec une de ses techniques: ***la Programmation Logique Inductive (PLI)***

2. Introduction à la Programmation Logique Inductive (PLI)

Programmation Logique Inductive (PLI)

- La Programmation Logique Inductive (PLI-*Inductive Logic programming*) est une **technique d'apprentissage symbolique** (Muggleton, 1991)(Lavrack&Dzeroski, 1994)
- En apprentissage supervisé, la PLI utilise **les clauses du premier ordre** pour obtenir une **représentation expressive uniforme** des exemples, de la base de connaissances et des hypothèses (règles)
- Cette representation :
  - Est **plus expressive** que la représentation **attribut-valeur** (*propositionnelle*) des méthodes **d'apprentissage statistiques**
  - Permet un **traitement de la langue naturelle plus profond**
  - Permet une **intégration facile** et **naturelle** de **connaissances de domaine (ontologie, thésaurus, …)** au processus d'apprentissage
- La PLI est très **proche des ontologies et du Web sémantique**

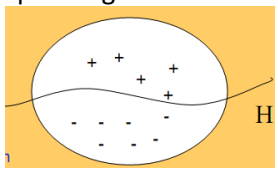
Apprentissage supervisé d'un classifieur basé sur la PLI

Entrée :

- **Base de connaissance ou "Background Knowledge" (BK)**
- Un ensemble **E** d'**exemples positifs** et **négatifs d'apprentissage**
- Une **étiquette de classification *c*** pour **chaque exemple d'apprentissage**

Sortie :

- Une théorie logique ou **Hypothèse H** séparant les exemples positifs des exemples négatifs



Apprentissage d'une règle en PLI

BK
Intentional:
- parent(X, Y) :- father(X,Y).
- parent(X, Y) :- mother(X,Y).

Extensional:
- father(pat,ann).
- father(tom,sue).
- female(ann).
- female(eve).
- female(sue).
- male(pat).
- male(tom).
- mother(eve,sue).
- mother(ann,tom).

Entrée

Examples:
Positive:
- daughter(sue,eve).
- daughter(ann,pat).

Negative:
- daughter(tom,ann).
- daughter(eve,ann).

Sortie **daughter(D,P) :- parent(P,D), female(D).**

Le prédicat **daughter** est induit à partir des exemples positifs et négatifs, ainsi que des prédicats **parent** et **female** déclarés dans la BK

Travaux reliés

La PLI est déjà utilisé en IE, principalement en RE:

- **(Seneviratne & Ranasinghe, 2011):** extraction d'une seule relation (located_in) sur un petit corpus de 13 pages de Wikipédia sur les oiseaux.
- **(Smole et al., 2011)**: apprentissage de règles pour extraire des informations à partir des définitions d'entités géographiques dans le texte (en langue slovène) pour l'extraction des 5 relations les plus fréquentes en 1308 définitions d'entités spatiales
- **(Kordjamshidi et al., 2012)**: « Spatial Role Labeling – SpRL", en combinant klog (P. Frasconi et al., 2014) enrironnement d'apprentissage relationnel basé sur les noyaux et un classificateur SVM.

Mais les corpus traités, le nombre de relations extraites et les performances sont limités.

3. Le système OntoILPER

OntoILPER

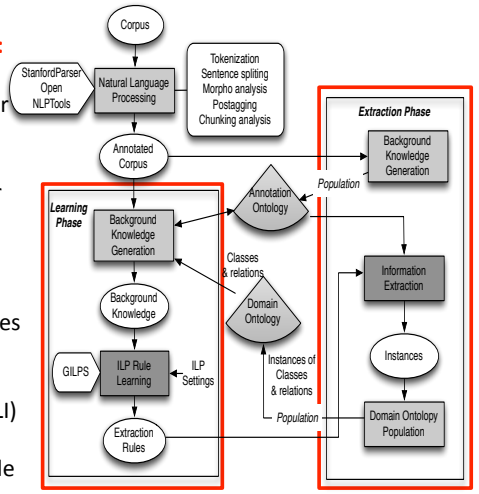
OntoILPER permet l'extraction **d'instances d'entités nommées (REN)** et de **relations binaires** (**RE**) de textes en ***anglais***

OntoILPER repose sur :

- Un **modèle relationnel des phrases** basé sur un **graphe des dépendances** (Marneffe&Manning, 2008), traitées comme des **prédicats logiques**
- Un **processus d'apprentissage basé sur la PLI**, induisant des ***règles d'extractions symboliques***
- Une **ontologie de domaine** :
  - **En entrée**: permet de ***choisir les concepts*** qui doivent être peuplés
  - **En sortie** : est peuplée par les instances extraites
- Une **ontologie d'annotation** :
  - **Permet de stocker les annotations**
  - **Utilisée pour appliquer les règles d'extraction**

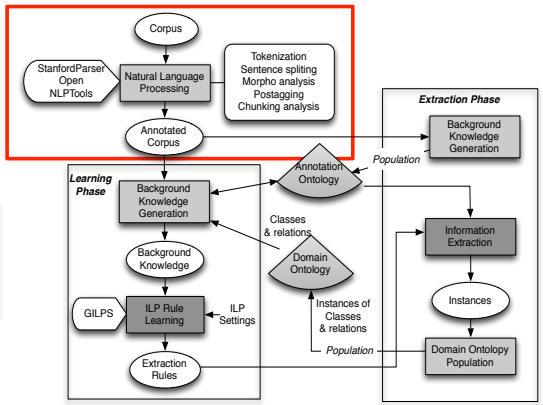
Architecture d'OntoILPER

- **2 phases:**
  - **PH1-Phase d'apprentissage:** Induction de règles symboliques d'extraction par PLI
  - **PH2- Phase d'extraction:** Extraction d'information par application de ces règles
- **Composants majeurs:**
  - TAL
  - Génération des connaissances de base (BK – Background Knowledge)
  - Apprentissage des règles (PLI)
  - Application des règles
  - Peuplement de l'ontologie de domaine



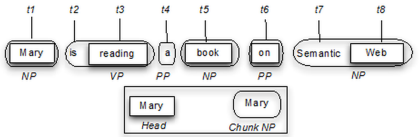
PH1 & PH2: Composant TAL (1)

Composant de traitement automatique du langage naturel (TAL)



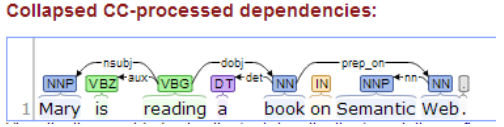
PH1 & PH2 : Composant TAL (2)

Chunking Analysis



Ce component utilise les outils Stanford CoreNLP & Open NLP

Dependency Graph



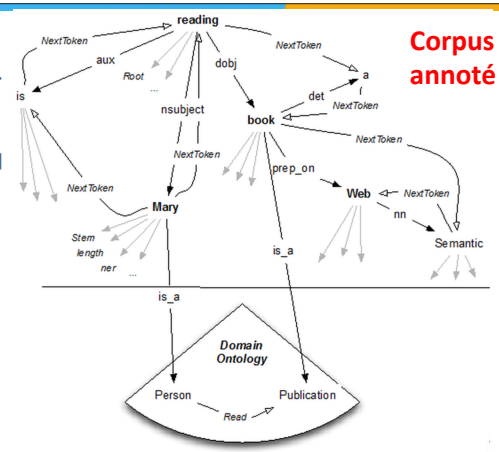
Il réalise: séparation de phrases, tokenisation, étiquetage morphosyntaxique (POS tagging, lemmatisation, analyse de chunks), extraction d'entités nommées (NER) et analyse des dépendances.

PH1 & PH2: Composant TAL (3)

Final Graph-based Model of sentences

Ground Predicates in BK

nsubj (reading, Mary)
aux (reading, is)
det (book, a)
head (Mary, NP)
nextToken (Mary, is)
nextToken (is, reading)
length (Mary, 4)
ner(Marry, person)

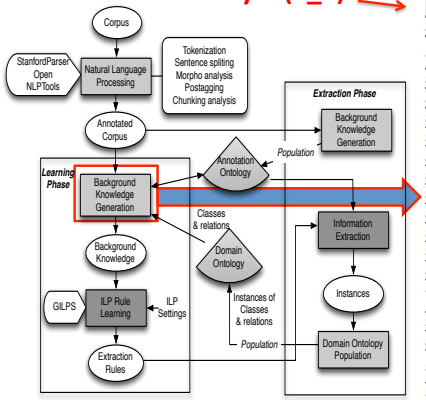


Corpus annoté

Phrase: « Mary is reading a book on Semantic Web »

PH1: Génération des connaissances de base (BK)

Pour le token « Mary » (T_1)



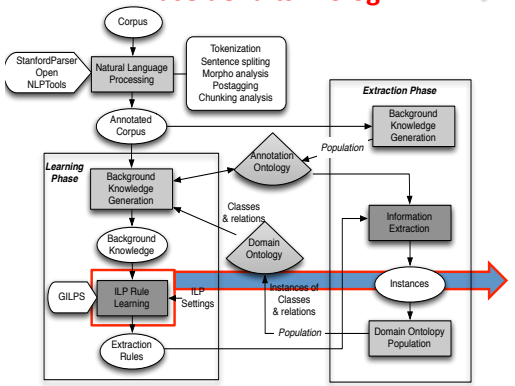
PROLOG PREDICATES FOR THE TOKEN "MARY" (T_1)

Predicates Generated	Meaning
token (t_1)	t_1 is the token identifier
t_dep (nsubj, t_3, t_1)	there is a noun subject dependency between token t_3 and t_1
t_next (t_1, t_2)	token t_2 follows token t_1
t_stem (t_1, "Mary")	the stemming of the token t_1 is "Mary"
t_length (t_1, 4)	t_1 has length of 4
t_orth (t_1, upperInitial)	t_1 has an initial uppercase letter
t_type(t_1, word)	t_1 is a word
t_pos (t_1, nnp)	t_1 is a singular proper noun
t_ner (t_1, person)	t_1 is a person entity
t_root (t_3)	t_3 is the root of the dependency graph
t_bigposbef (t_n, ...)	POS tag bigram of the tokens after t_n
t_bigposaft (t_1, vbz-vbg)	POS tag bigram of the tokens before t_1
t_trigposbef (t_n, ...)	POS tag trigram of the tokens before t_n
t_trigposaft (t_1, vbz-vbg-dt)	POS tag trigram of the tokens after t_1
t_isHeadNP (t_1)	t_1 is the head of the nominal chunking
t_isHeardVP (t_n...)	t_n is the head of the verbal chunking
t_isHeardPP (t_n....)	t_n is the head of the prepositional chunking
t_ck_tag (t_1, NP)	t_1 is part of a nominal chunking

Base de faits Prolog

PH1: Composant PLI d'apprentissage de règles

Base de faits Prolog



Utilise de système noyau de PLI « GILPS » (Santos 2010)

```
% MODE DECLARATIONS
:-modeh( 1, located_in(+token, +token) ).

:-modeb(*, t_hasDep(#dep, +token, -token)).        Relational features
:-modeb(1, t_next(+token, -token)).

:-modeb(1, t_root(+token)).                        Morpho syntactical features
:-modeb(1, t_stem(+token,#string)).
:-modeb(1, t_pos(+token,#postag)).
:-modeb(1, t_length(+token,#int)).
:-modeb(1, t_orth(+token,#orth)).
:-modeb(1, t_type(+token,#type)).
:-modeb(1, t_ck_ot(+token,#ck_tag)).
:-modeb(1, t_ner(+token,#ner)).
:-modeb(1, t_ck_tag_ot(+token,#string)).
:-modeb(1, t_gpos(+token,#gpos)).

:-modeb(1, t_isHeadNP(+token)).                    Chunking features
:-modeb(1, t_isHeadVP(+token)).
:-modeb(1, t_isHeadPP(+token)).

:-modeb(1, t_bigPosBef(+token,#bigposbef)).        N-grams features
:-modeb(1, t_bigPosAft(+token,#bigposaft)).
:-modeb(1, t_trigPosBef(+token,#trigposbef)).
:-modeb(1, t_trigPosAft(+token,#trigposaft)).
```


Exemples de règles d'extraction induites

Règle pour la relation "located_in":

- **located_in (A,B):- t_class(A, loc), t_next(A, B), t_class(B, loc).**
  → Cette règle caractérise le patron "City, Country", ex. "Marseille, France"
- **located_in (A, C) :- t_next(A,B), t_next(B, C), t_ner(A, org), t_class(C, loc).**
  → Cette règle caractérise le patron "ORG, [at|in|on] LOC", ex. "White House in USA"

…

Règles pour d'autres relations:

- **part_whole (A,B):-** t_gpos(A, nn), t_next(A, B), t_subtype(B, state-or-province).
- **part_w(A,B):-** t_next(A,B), t_pos(A,nnp), t_ne_type(B,gpl), t_subtype(A,pop-center).
- **live_in(A,B)**:- t_pos(A,nn), t_class(A, person), t_hasDep(amod,B,C), t_next(C,B), t_class(B, loc), t_isHeadNP(B).

Ces règles sont compréhensibles (et modifiables) par des humains …

4. Evaluation Expérimentale

Protocole expérimental

- **2 corpus de références : reACE 2004/2005 datasets (broadcast news):**

Relation Type & Subtype

reACE 2004 - Relation Type/Subtype Hierarchy	Freq
Employee-Membership-Subsidiary (EMP_ORG)	
Employee-Staff	303
Employee-Executive	220
Member-of-Group	80
General-Affiliation (GEN_AFF)	
Located	352
Citizen-Resident-Religion-Ethnic	98
Part-Whole (PRT_WHOLE)	
Part-Whole	174
Subsidiary	100
Personal-Social (PER_SOC)	
Business	35
Family	15
Total	**1377**

reACE 2005 - Relation Type/Subtype Hierarchy	Freq
Organization-Affiliation (ORG_AFF)	
Employment	228
Membership	36
General-Affiliation (GEN_AFF)	
Located	280
Citizen-Resident-Religion-Ethnic	39
Part-Whole (PRT_WHOLE)	
Geographical	119
Subsidiary	47
Personal-Social (PER_SOC)	
Business	16
Family	42
Total	**807**

- **Métriques d'évaluation : Precision (P), Recall (R) et F1-measure (F1)**
- **Theory Compression Ratio** *(mesure de généralité de la règle pour éviter le sur-apprentissage:*

$$TCR = \frac{\text{Nb de règles dans la théorie apprise}}{\text{nb d'exemples positifs dans l'ensemble d'apprentissage}}$$

- **5 validations croisées**

Contribution des "features":

Quelle combinaison de BK linguistique est la meilleure ?

ID	Features	reACE 2004 P	R	F1	reACE 2005 P	R	F1
1	Baseline	81.09	39.81	53.40	60.53	25.12	35.52
2	+C	80.17	47.13	59.36	75.05	34.03	46.80
3	+D	81.01	46.93	59.43	72.91	36.51	48.65
4	+D+C	89.01	54.40	67.53	74.81	38.14	50.48
5	+D+C+P	91.16	62.04	73.83	81.75	44.24	57.37
6	+D+C+P+Cr	93.30	66.68	77.77	83.68	50.43	62.91
7	+D+C+P+Cr+N	93.04	67.12	77.99	80.59	51.39	62.68
8	+D+C+P+Cr+A	92.20	71.13	80.31	83.03	63.38	**71.86**
9	+D+C+P+Cr+A+N	92.91	73.07	**81.80**	82.30	61.85	70.62

P = Precision
R = Rappel
F1= F1-measure

Relation subtypes

Baseline = Morphological + Next
C = Nominal and verbal chunkings
D = Dependencies
→ Structural features

P = POS tagging
Cr = Chunking-related features
→ Attributive features

N = NER
A = reACE Corpus types
→ Semantic features

Résultats de classification sur les sous-types de relations et les relations-types (1)

Performance results of relation subtypes on both datasets

Relation subtypes

Rel. Type	reACE 2004 Rel. Subtype	P	R	F1	reACE 2005 Rel. Subtype	P	R	F1
EMP_ORG	Employ-Staff	78.10	86.90	82.27	Employ	89.60	86.22	87.88
	Employ-Exec	95.49	77.00	85.25	-	-	-	-
	Member	92.18	76.82	83.80	Member	94.30	71.03	81.03
GEN_AFF	Citizen-Resident	98.81	69.58	81.66	Citizen-Resident	100.00	61.10	75.85
	Located	83.28	80.09	81.65	Located	86.00	84.10	85.04
PERS_SOC	Business	100.00	69.42	81.95	Business	0.00	0.00	0.00
	Family	100.00	39.11	56.23	Family	92.70	57.70	71.13
PRT_WHL	Part-Whole	93.20	83.38	88.02	Geo	100.00	62.10	76.62
	Subsidiary	95.10	75.30	84.05	Subsidiary	95.80	72.51	82.54
	Avg	**92.91**	**73.07**	**81.80**	**Avg**	**82.30**	**61.85**	**70.62**

Il est plus difficile de classifier à un niveau détaillé (sous-type de relation)

Relation types

Classification results of relation types on the reACE 2004

Rel. Type	#E+	#Rules	TCR	P	R	F1
EMP_ORG	603	65	0.11	86.00	84.00	84.99
GEN_AFF	450	51	0.11	86.90	78.90	82.71
PER_SOC	50	18	0.36	100.00	64.40	78.35
PRT_WHL	274	33	0.12	91.00	81.60	86.04
Total	1377	167				
Avg			0.18	90.98	77.23	83.54

Classification results of relation types on the reACE 2005

Rel. Type	#E+	#Rules	TCR	P	R	F1
ORG_AFF	264	38	0.14	88.70	77.80	82.89
GEN_AFF	319	60	0.19	94.40	70.40	80.65
PER_SOC	58	19	0.33	100.00	58.30	73.66
PRT_WHL	166	35	0.21	87.60	72.30	79.22
Total	807	152				
Avg			0.22	92.68	69.70	79.56

Résultats de classification sur les sous-types de relations et les relations-types (2)

Classification de types/sous-types relations :

- Selon (Zhou et al., 2005/2007), il **est plus difficile de classifier des relations à des niveaux plus profonds** de la hiérarchie **parce que :**
  - **il y a moins d'exemples par classe**,
  - les **classes deviennent plus semblables** à mesure que le niveau de classification devient plus profond,

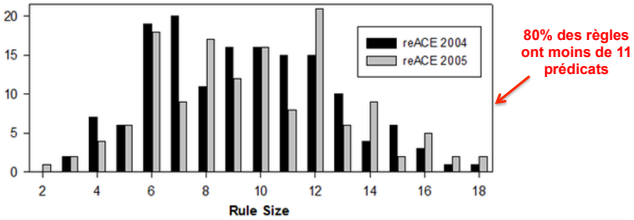
Theory Compression Ratio (TCR):

- **Plus de règles nécessaires pour couvrir les exemples** du corpus **reACE 2005**
- Le **TCR moyen est de 0.22 for reACE 2005** et de **0.18 pour reACE 2004.**
- *Cette tendance des scores TCR se retrouve aussi dans la mesure F globale, nettement inférieure pour l'ensemble de données reACE 2005*
- **La raison pourrait être :** présence d'exemples plus **complexes** dans **reACE 2005** seulement couverts par :
  - **Des règles plus longues** ou
  - **Des règles avec un degré de généralisation plus bas**

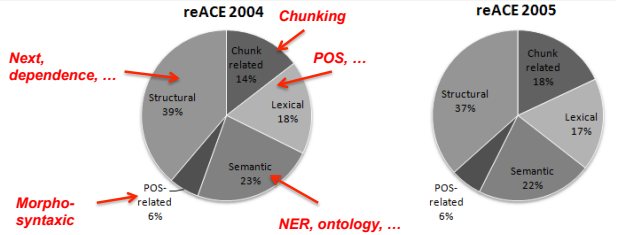
Analyse des règles induites :

Aspects qualitatifs des règles apprises

Distribution de la taille des règles



Ratio des types des prédicats dans l'ensemble des règles induites



5. Conclusion et Perspectives

Conclusion

OntoILPER est un système d'extraction de relations basé sur la PLI:

- S'appuyant sur un **riche modèle graphique des phrases**
- Il exploite une **ontologie de domaine** en **entrée** et **en sortie** du processus d'extraction, ainsi que d'une **ontologie auxilliaire** (d'annotation linguistique)
- **Son composant d'apprentissage basé sur la PLI** induit des **règles d'extraction symboliques** compréhensibles par des humains à partir **d'exemples annotés**
- Il permet **d'intégrer de la connaissance utile dans la BK** pour amméliorer la **qualité des règles d'extraction induites**
- **Les expérimentation montrent que OntoILPER est performant dans 2 domaines** : les News (reACE datasets) et en Biomedical (LLL and IEPA datasets)

Perspectives

Optimisation :

- Utilisation de **methodes ensemble**, pour améliorer les performances du **processus d'apprentissage (PH1)**
- Utilisation d'un **triple-store** et des **techniques du Web Sémantique** pour améliorer l'application des règles symboliques dans le **processus d'extraction (PH2)**

Performances :

- **D'intégrer plus de connaissances dans la BK** au niveau du prétraitement, pour prendre en compte les **aspects sémantiques** (prédicats sur les synonymes, hypernyms / hyponymes, rôles sémantiques ...)
- **Extraction d'évènements** : Actuellement OntoILPER extrait uniquement des relations binaires, nous projetons d'étendre OntoILPER pour **extraire des relations n-ary (Event Extraction).**

Publications associées

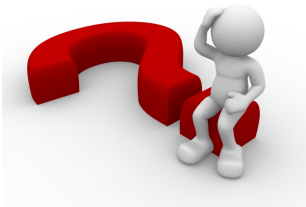
Revues :

- R. Lima, B. Espinasse, F. Freitas (2017), « Extraction automatique d'entités et de relations par ontologies et programmation logique inductive », in: Knowledge and Information System Journal (**KAIS**), 2017 or 2018 (***Accepté pour publication***).
- B. Espinasse, R. Lima, F. Freitas (2016), « Extraction automatique d'entités et de relations par ontologies et programmation logique inductive », in: Revue d'Intelligence Artificielle (**RIA**), Vol. 30 (n° 6/2016), dec 2016 (Répertoriée Scopus et DBLP).

Conférences :

- B. Espinasse, Lima R., Magdy D., « Extraction automatique d'entités et de relations par ontologies et programmation logique inductive » , Journée Francophones sur les Ontologies - **JFO 2016**, 13-14 Octobre 2016, Bordeaux, France..
- R. Lima, S. B. Espinasse, F. Freitas « Relation Extraction from Texts with Symbolic Rules Induced by Inductive Logic Programming », IEEE International Conference on Tools with Artificial Intelligence, **IEEE-ICTAI 2015**, Vietri sul Mar, Italy, 9-11 nov. 2015.
- R. Lima, B. Espinasse, H. Oliveira, F. Freitas « Ontology Population from the Web: an Inductive Logic Programming-Based Approach », 11th International Conference on Information Technology: New Generations, **ITNG 2014**, Las Vegas, Nevada, USA, April 7-9, 2014.
- R. Lima, B. Espinasse, H. Oliveira, L. Pentagrossa, F. Freitas, « Information Extraction from the Web: An Ontology–Based Method using Inductive Logic Programming », IEEE International Conference on Tools with Artificial Intelligence, **IEEE-ICTAI 2013**, Washington DC, USA, November 4-6, 2013.
- R. Lima, B. Espinasse, H. Oliveira, R. Ferreira, L. Cabral, F. Freitas, R. Gadelha, « An Inductive Logic Programming-Based Approach for Ontology Population from the Web », **DEXA 2013**, Prague, Czech Republic, August 26-29, 2013.

Merci de votre attention, Avez-vous des questions ?



Evaluation comparative : Corpus Biomedical

Cross-validation results of the RE systems on the PPI corpora

Corpus	OntoILPER		Miwa *et al.* (2010)		Quian/Zhou (2012)		Tikk *et al.* (2010)		Airola *et al.* (2008)	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
LLL	79.9	85.2	82.9	*90.5*	**84.6**	89.9	79.1	86.8	76.8	83.4
HPRD50	**75.3**	*87.6*	75.0	86.6	68.8	83.7	69.7	84.0	63.4	79.7
IEPA	76.1	87.2	**77.8**	*88.7*	69.8	82.8	70.7	81.0	75.1	85.1

- Comparative evaluation with statistical-based RE systems over the Protein-Protein Interaction corpora (biomedical domain)
- The selected RE systems are kernel methods (Support Vector Machines) [Miwa et al., 2010] [Quian & Zhou, 2012] [Tikk et al., 2010] [Airola et al., 2008].
- **OntoILPER obtained very competitive results**