RelTEX : Usage du corpus ISTEX pour l'exploitation de méthodes d'extraction de connaissances à partir de textes

Nathalie Aussenac-Gilles aussenac@irit.fr
Farah Bénamara - Catherine Comparot
Mouna Kamel - Cassia Trojahn
Elena Manishina (Post-doc à partir du 1/07/2016)
Collaboration avec Cécile Fabre et Anne Condamines (CLLE-ERSS)

RelTEX : projet chantier d'usage

- Domaines scientifiques concernés
  - TAL, extraction de connaissances
  - Web sémantique
- Problématique
  - **Usage du corpus ISTEX pour l'exploitation de méthodes d'extraction de connaissances à partir de textes**
- Problématique plus précise
  - Extraction de relations sémantiques entre entités ou entre classes à partir des textes
  - Évaluation d'approches faisant appel au langage naturel et à la structure des textes
- Projet complémentaire SemPédia (2015-2018)
  - financé par la région Midi-Pyrénées et la COMUE de Toulouse
  - Plateforme d'extracteurs de relations pour Wikipedia en français

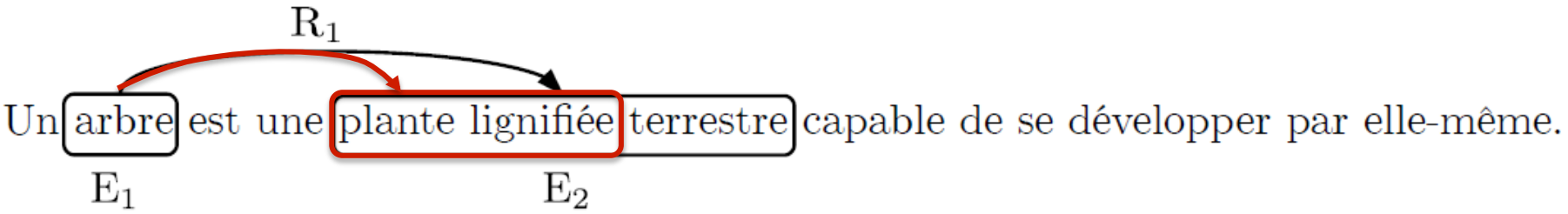
Plan de la présentation

- Extraction de relations sémantiques
  - Qu'est-ce qu'on cherche ?
  - Comment trouver des relations ?
- Le projet RelTEx
  - Projet envisagé
  - Questions préalables
  - Pistes

Extraction de Relations

- Tâche appartenant à l'extraction d'information (IE) et qui vise à extraire des relations sémantiques entre des entités du monde.
- Tâche cruciale pour la construction de ressources ou applications TAL :
  - e.g. : WordNet, DBpedia, etc.
  - e.g. : Recherche d'Information, Systèmes Question-Réponse, etc.

- **Exemple** : relation R_1 d'hyperonymie entre E_1 et E_2 :



- Une fois extraite, cette information est formalisée (FOL, DL, etc.) :
  e.g. : $\forall x(\text{arbre}(x) \implies \text{plante-lignifiée-terrestre}(x))$
  e.g. : $\textit{arbre} \sqsubseteq \text{plante-lignifiée-terrestre}$

Qu'entend-on par relation sémantique ?

Domaine de recherche

- Linguistiques: traces de relations en discours

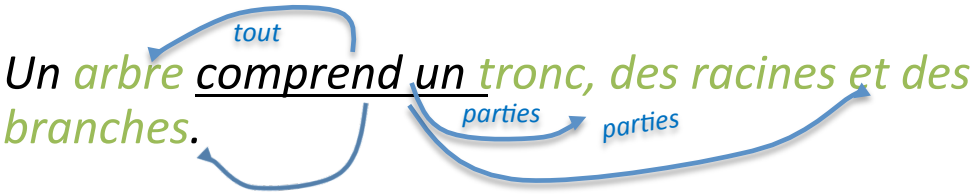
  Frame

- Terminologie
  - Structure hiérarchique
  - Modélisation en BD ou SKOS

- Extraction d'information
  - Classes et relations connues
  - Alimenter une BD
  - Relations entre instances

Ce qu'est une relation sémantique

Un arbre est une plante lignifiée terrestre capable de se developer par elle-même.



Un arbre [arbre] est une [hyperonymy] plante lignifiée [plante lignifiée] terrestre [propriété] capable de se developer par elle-même.

Arbre *est_un* plante lignifiée, Arbre *AEcosystème* Terrestre

tree	Plantation year	Species	Branches
Tree1	1990	Oak	> 20
Tree2	1995	Oak	15

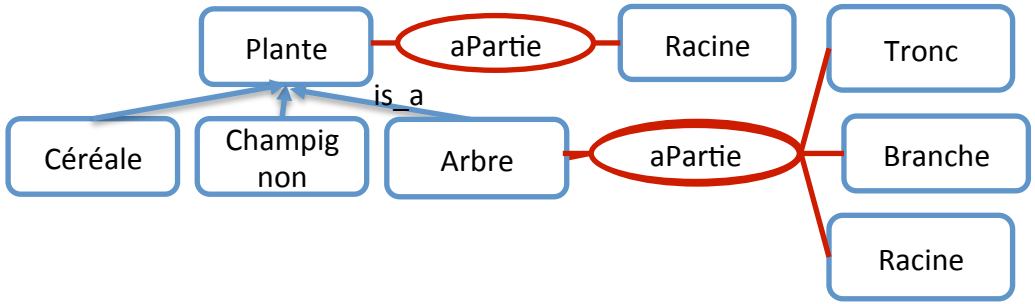
Qu'entend-on par relation sémantique ?

Domaine de recherche

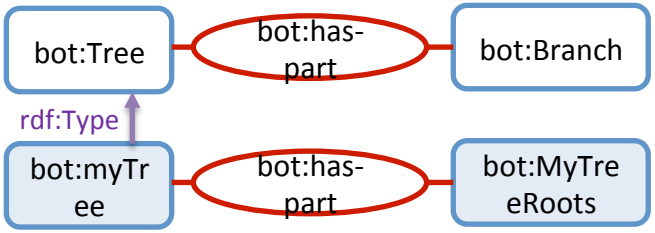
- Construction d'ontologie : relation sémantique
  - Formalisée (logique, RDF, OWL …)
  - Permet le raisonnement
  - fait partie d'un graphe

- Web semantique
  - Triplet indépendant
  - Publié
  - Format RDF
  - Lié à d'autres données

Ce qu'est une relation sémantique



bot:Tree bot:has_part bot:Branch



Extraction de relations

Trois grandes familles de méthodes :

1. Approches symboliques,
2. Approches statistiques,
   - supervisées,
   - semi-supervisées,
   - non supervisées.
3. Approches mixtes.

Exploiter les régularités d'expression des relations en langue

-> on sait quels types de relation on cherche

-> on apprend les types de relations en corpus (open relation extraction)

Facteurs influençant leur choix :

- évolution technologique (capacités de traitement, stockage, algorithmes),
- domaine visé (spécialisé ou non),
- genre du corpus (journalistique, encyclopédique, etc.),
- nature des sources (structurées, non structurées, etc.),
- visée applicative (construction de ressources, etc.),
- utilisation de ressources (ressources distributionnelles, thésaurus, etc.).

Extraction de relations

Approche symbolique : patrons

Intuition : intégrer manuellement des connaissances linguistiques.

- Travaux : (hyperonymie) [Hearst, 1992], (méronymie) [Berland and Charniak, 1999], (multiples) [Aussenac-Gilles and Jacques, 2008]
- Exemples de patrons :
  - Y tel que X
  - X et/ou autres Y
  - Y incluant X
  - X est une sorte de Y

Patrons incluant la recherche des arguments

- Det NP tel que Det NP
- NP_plur et/ou autres NP_plur
- NP incluant Det NP
- NP est une sorte de NP qui …

is-a(sérogroupage, test-de-présence-bactérienne)

Des **tests de présence bactérienne** tels que le **sérogroupage** peuvent être utilisés dans certains cas.

is-a(voiture, véhicule-motorisé)

Les cars, les motos, les **voitures** et autres **véhicules motorisés** sont interdits dans l'enceinte du bâtiment.

Extraction de relations

Approches statistiques : supervisées

Intuition : intégrer de manière automatique les connaissances linguistiques au moyen de données annotées.

- Travaux : (svm) [Zelenko et al., 2003], (maxent) [Kambhatla, 2004], (neural network) [Rosario and Hearst, 2004]

- Deux sous-tâches :
  - Décider si 2 entités sont reliées,
  - Identifier la nature de la relation.

- Traits : contextes des entités, span entre les entités, types des entités, séquence des chunks, arbre de constituants et/ou de dépendances, etc.

est-pdg-de(Guillaume Faury, Eurocopter)

Eurocopter prendra un nouveau nom à partir de janvier 2014 dans le cadre d'une restructuration, a déclaré le PDG **Guillaume Faury**.

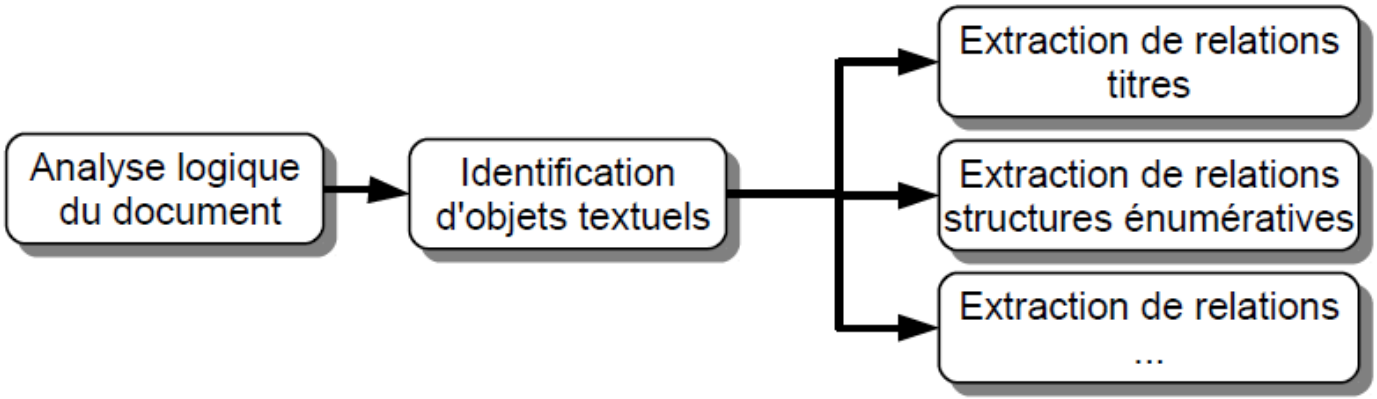
Travaux menés au sein de MELODI

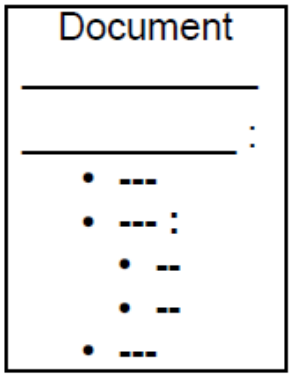
- Identification de la structure logique (profonde) de documents
- Patrons de différents types de relations (Jacques et Aussenac, 2005)
- Recherche de relations sur plusieurs phrases
- Prise en compte de la structure logique comme indice supplémentaire
  - Structure obtenue à partir de la mise en forme
  - Analyse au niveau discours, utilisation de la SRT
  - Corpus Wikipedia

Apprendre des relations en utilisant la structure des documents – Thèse J.P. Fauconnier

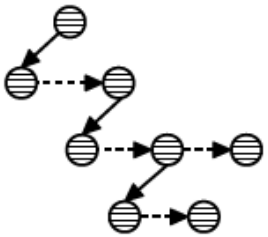
Approche proposée :

- Objectif : extraction de relations en utilisant la structure des documents,
- Intuition : adapter le traitement selon nature/rôle/position d'un objet textuel dans la structure du document.

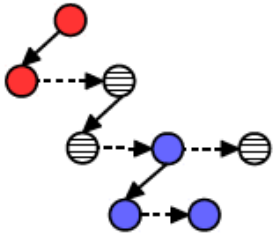




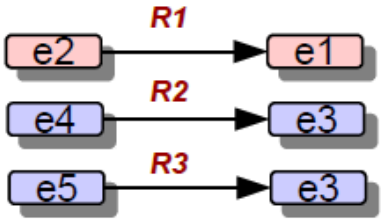
Arbre de dépendance



Identification d'objets textuels



Extraction de relations



Application aux structures énumératives

Th. J.P. Fauconnier

IS_A

Types de chaussures [modifier | modifier le code]

Chaussures classiques [modifier | modifier le code]

Certaines chaussures sont portées exclusivement par les femmes :

- Escarpin, parfois nommé décolleté (il existe aussi des escarpins pour homme mais leur usage est tombé en désuétude)
- Salomé
- *Stiletto* ou talon aiguille
- Ballerine
- Découpé

IS_A

Les chaussures portées uniquement par les hommes sont le derby et le richelieu. Il existe maintenant des derbys et des richel adaptées avec un talon.

Le mocassin est porté aussi bien par les hommes et que par les femmes. Les variantes du mocassin sont le loafer (parfois no loafer. Les chaussures dont la tige recouvre la jambe sont la botte et ses variantes : la bottine, le bottillon, et la cuissarde.

Chaussures légères et d'intérieur [modifier | modifier le code]

Certaines chaussures ont des tiges très légères, comme :

- Sandale
- Sandalette
- Nu-pied
- Tongs
- Mule

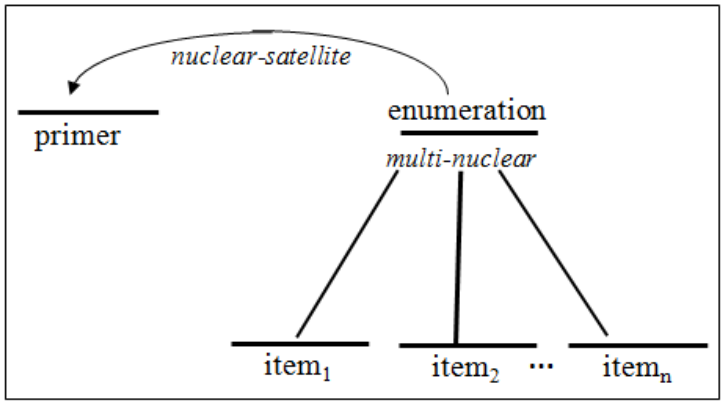
Les chaussures portées en intérieur sont :

- Pantoufle
- Charentaise
- Chausson

Autres types de chaussures [modifier | modifier le code]

Les rangers sont des chaussures en cuir, à long lacets, montant jusqu'aux chevilles ou plus haut. Elles sont utilisées par les armées du monde entier pou chaussures sont conçues pour protéger les pieds contre les risques de nature électrique, chimique, mécanique, thermique (voir : chaussure de sécurité)

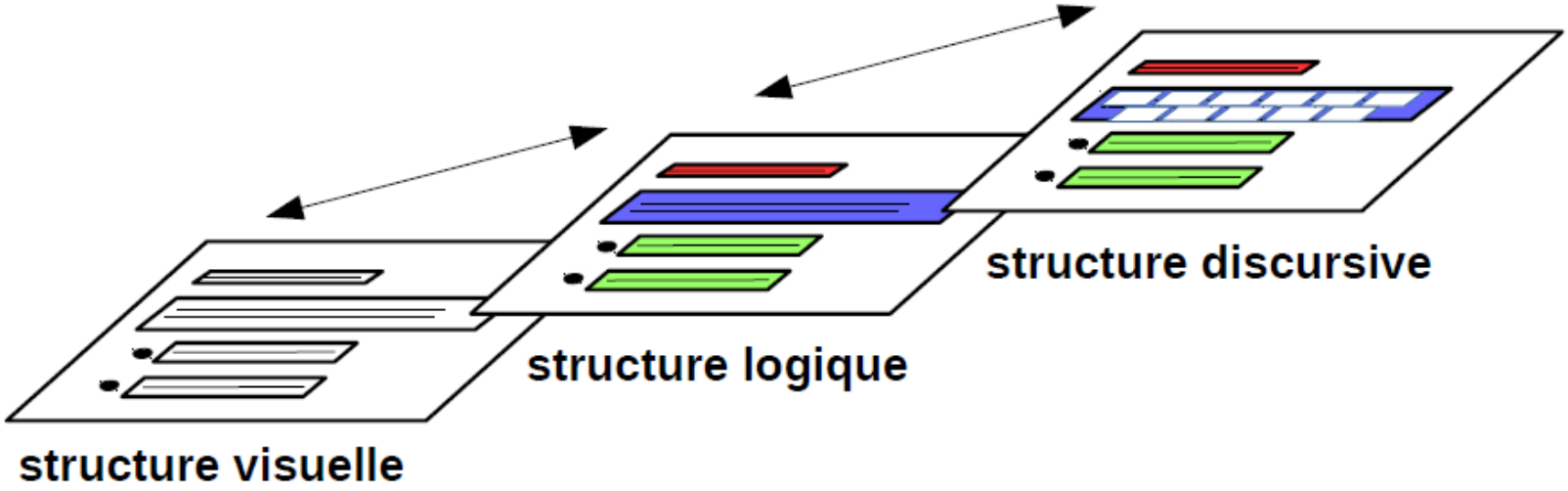
Les chaussures utilisées en danse sont :



Apprentissage de structures énumératives parallèles

- Tâche de classification pour identifier la nature de la relation (est_un, partie_de, autre
- Extraction de termes pour identifier les arguments dans l'amorce et les items

Trois niveaux de structure du document :



Segmentation en unités :

Consensus « flou » dans la littérature

- **Structure visuelle** : alinéa et blocs visuel, (//OCR)
- **Structure logique** : titre, paragraphe, item, etc. (//HTML, LaTeX)
- **Structure discursive** : EDU et CDU.

La structure logique profonde → un arbre de dépendance :

- Deux relations : subordination et coordination,
- Principe de dépendance : unités liées dans la cohérence du document,
- Composants : nœuds = unités élémentaires, arcs = dépendances typées.

document

1. Titre

P_____

P_____

2. Titre

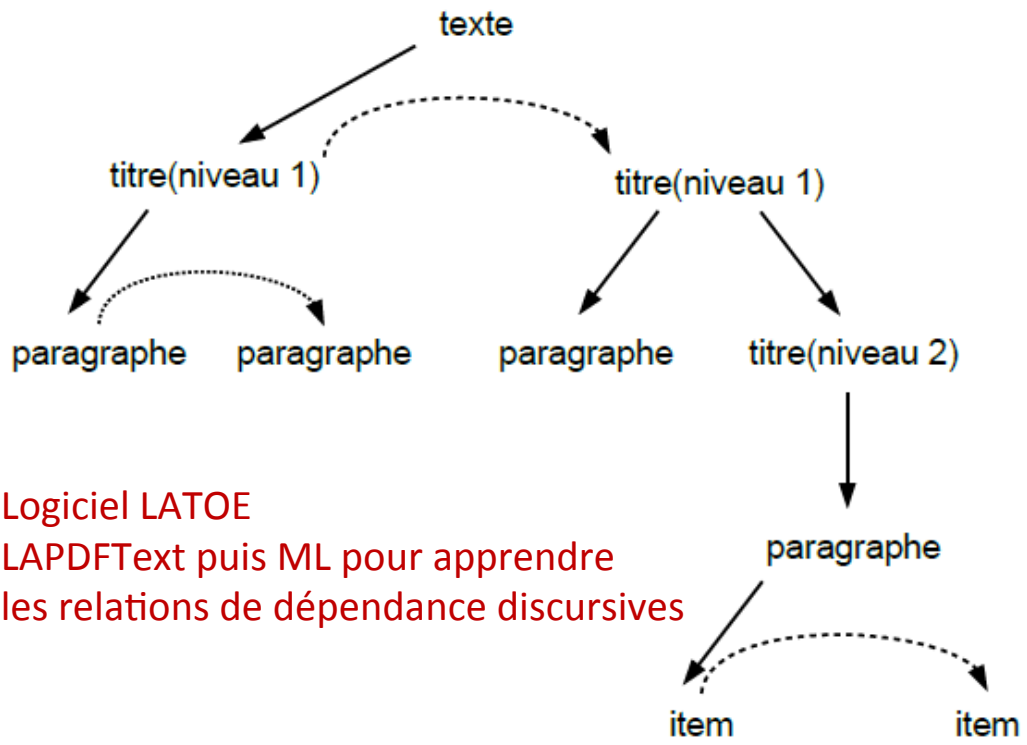
P_____

2.1. Sous-titre

P_____

- item_____ ;
- item_____ ;

structure logique profonde



Logiciel LATOE
LAPDFText puis ML pour apprendre
les relations de dépendance discursives

Apprendre des relations à partir de structures énumératives - Th. J.P. Fauconnier

General Features	Description
POS	The presence of a part of speech in the primer or in the items
Start/End	The first or last part of speech in the primer or in the items
Plural	Boolean indicating the presence of a plural noun
Form	The number of tokens and the number of sentences
Primer's features	
Marker	Boolean indicating the presence of a relational marker
Syntactic	Boolean indicating if the last sentence is not syntactically complete, i.e. it ends with a subordinating conjunction, a preposition, a verb, etc.
Punctuation	Returns the last punctuation

- Corpus
  - 745 structures énumeratives de pages Wikipedia
  - 3 types de relation : taxonomique, ontologique_non_taxonomique, non_ontologique
- Tâche de Classification
  - Choix des traits
  - Evaluation automatique des traits
  - Comparaison de 3 algorithmes : SVM, MaxEntropy et reference (majorité)
  - Entraînement de 2 algorithmes
- Résultats
  - 82% f-measure pour SVM
  - Meilleurs résultats en séparant 2 étapes chacune avec ses traits (ontologique O/N puis taxonomique O/N)

RelTEX

- À court terme : étendre l'approche à des corpus scientifiques
  - Focus sur peu de types de relations
    - Hypéronymie, méronymie
  - Proposer plusieurs techniques complémentaires
    - Patrons, composition de mots, co-occurrence
  - Exploiter des indices linguistiques divers
    - Lexique, syntaxe, dépendances, relations discursives et structure logique des documents
- Vers une plateforme d'extraction de relations
  - Proposer plusieurs extracteurs complémentaires
    - Techniques différentes
    - Types de textes, domaines spécifiques
  - Traiter de nouveaux types de relations
  - Pouvoir ajouter des extracteurs selon les besoins

Atouts de la plateforme ISTEX

- La collection ISTEX
  - certains documents disponibles en PDF et XML
  - Volume -> approches par apprentissage
  - Articles scientifiques structurés
- Services disponibles
  - Grobid extraction de la structure
  - TermSuite Extraction de termes > arguments des relations entre classes
  - Unitex + CasSys Extraction des EN > arguments des relations entre entités ou typage des entités

RelTEX : ajustement du projet

- XML dans ISTEX
  - Riches méta-données
  - Texte simple entre balises <BODY> </BODY>
- XML pour expliciter la structure logique
  - Balises riches et hiérarchisées (titre1, titre2, …)
  - identifiables à partir du pdf (GroBid, LATOE)
- Nature des textes : Articles scientifiques
  - Peu de définitions, moins de relations d'hyperonymie
  - Relations de domaine
  - Quelques énumérations verticales
  - Identifier la sémantique de la structure par domaine / type de revue etc.

Questions préliminaires

- Techniques d'extraction
  - Patrons existants + Prise en compte de la structure
  - Apprentissage sur corpus annoté par une ressource
- Choix du corpus
  - Domaine : disponibilité d'une ressource sémantique ; termes et/ou EN déjà étiquetées > food
  - Langue(s) : F et GB
- Quel type(s) de relation ?
  - Interdomaine : est-un, partie_de
  - Spécifiques au domaine étudié
- Validation
  - Annotation automatique par la ressource
  - Comparaison à un "gold" ou à une ressource
  - Valider la nature de la relation vs valider les arguments

Premières études envisagées

- Choix domaine et corpus
  - Test de couverture du corpus par une ressource (corpus « food » ?)
- Évaluation de la présence de relations est-un sur échantillon > choix d'autre types de relations
- Tests des logiciels disponibles
- Elena Minishina débute au 1er juillet
  - Compétences TAL, structure logique et analyse de relations sémantiques

Pattern based relation extraction, an issue: variation

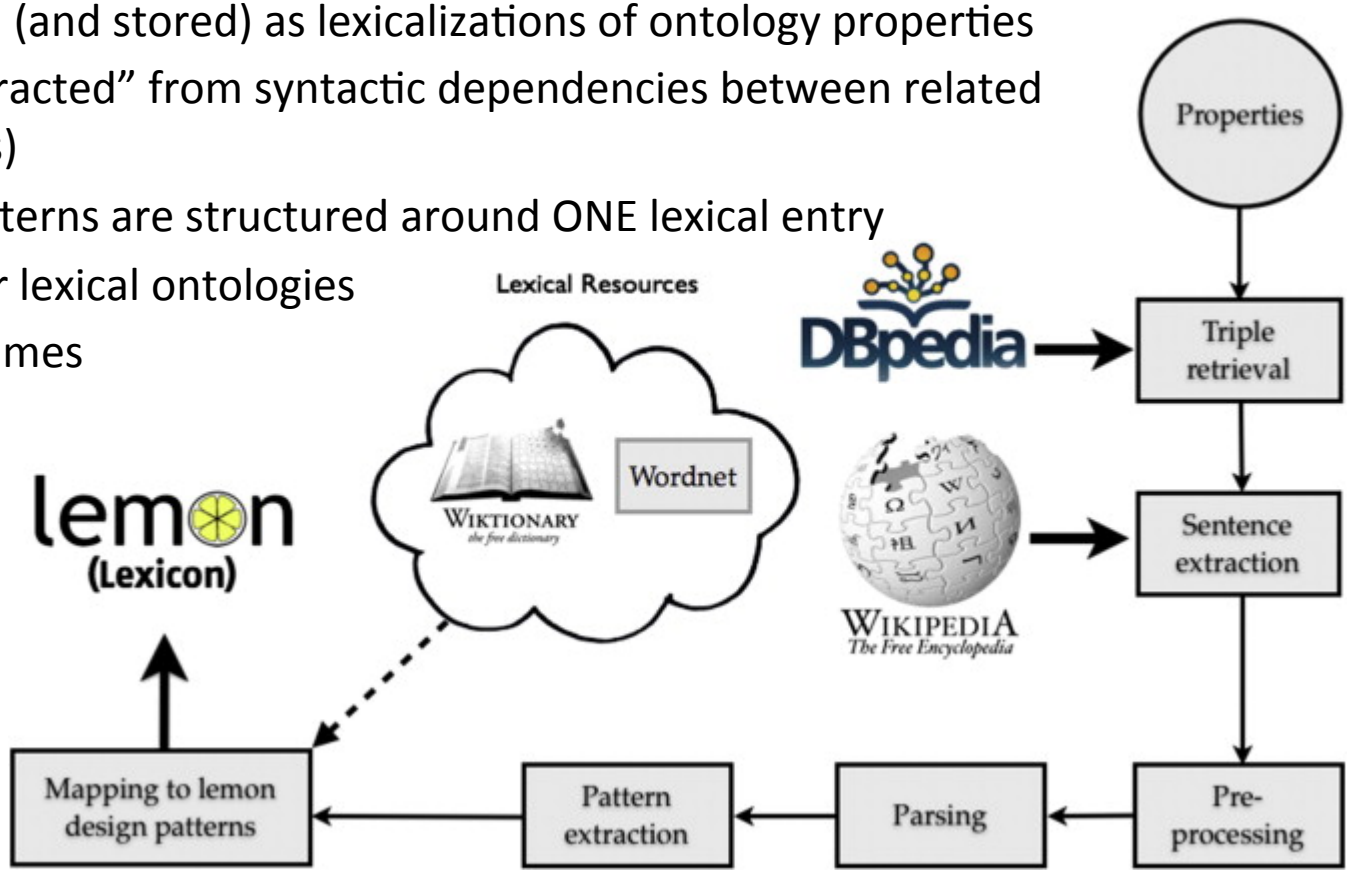
- *A tree comprises at least a trunk, roots and branches.*
- *With branches reacing the ground, the willow is an ornamental tree.*
- *The tree of the neighbor has been delimed.*
- *He climbs on the branches of the tree.*
- *This tree is wonderful. Its branches reach the ground.*

- Contains: very systematic pattern; the parts may be difficult to spot; enumeration > various parts
- With: meronymy pattern only in some genres (such as catalogs, biology documents)
- Delimed : Term and pattern are in the same word; requires background knowledge: *delimed* -> has_part branches (and branches are cut)
- Of : Very ambiguous pattern; polysemy reduced in [verb N1 of N2]
- Its : very ambiguous pattern; necessity to take into account two sentences

Pattern based relation extraction, learning patterns (1)

ATOLL—A framework for the automatic induction of ontology lexica
S. Walter, C. Unger, P. Cimiano, DKE (94), 148-162 (2014)

- Patterns are seen (and stored) as lexicalizations of ontology properties
- Patterns are "extracted" from syntactic dependencies between related entities (in triples)
- Assumes that patterns are structured around ONE lexical entry
- Lemon format for lexical ontologies
- Entries can be frames

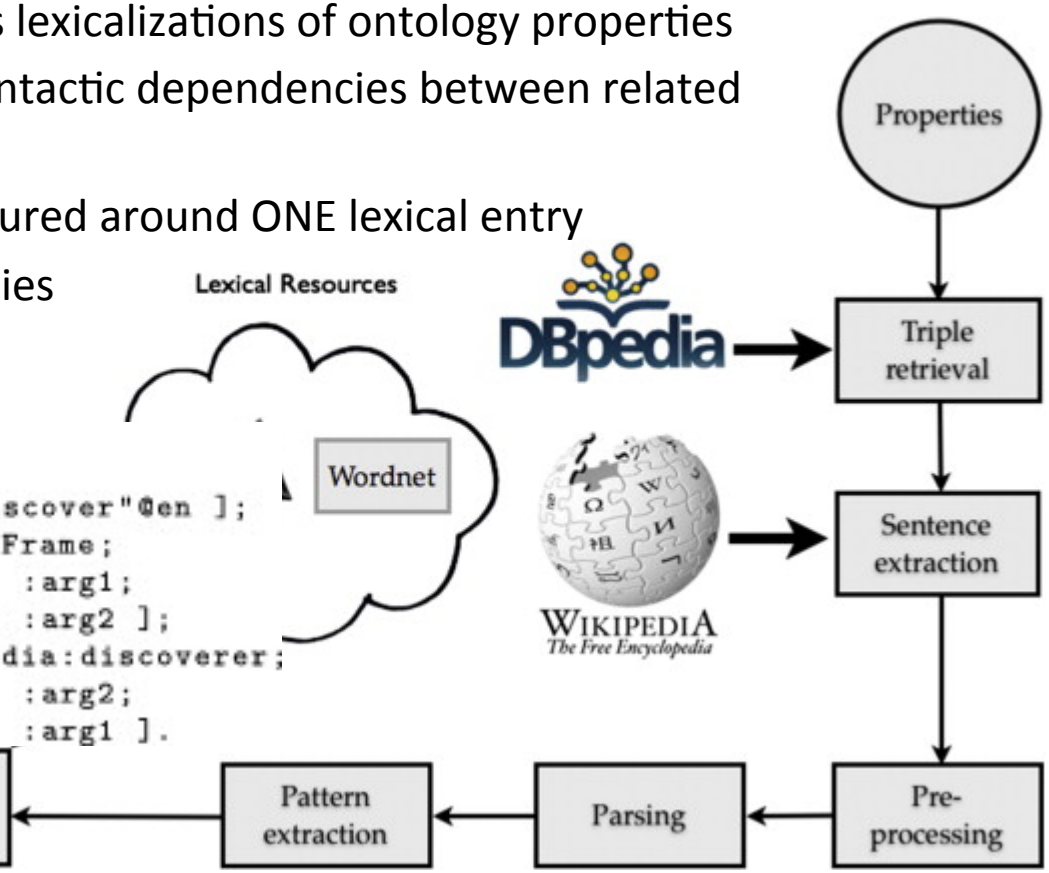


Pattern based relation extraction, learning patterns (1)

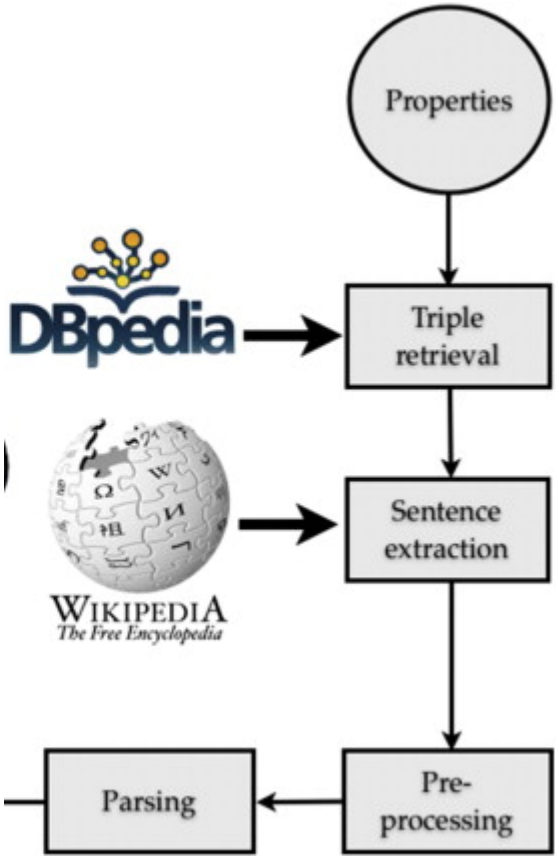
ATOLL—A framework for the automatic induction of ontology lexica
S. Walter, C. Unger, P. Cimiano, DKE (94), 148-162 (2014)

- Patterns are seen (and stored) as lexicalizations of ontology properties
- Patterns are “extracted” from syntactic dependencies between related entities (in triples)
- Assumes that patterns are structured around ONE lexical entry
- Lemon format for lexical ontologies
- Entries can be frames

```
:discover a lemon:Word;
 lexinfo:partOfSpeech lexinfo:verb;
 lemon:canonicalForm [ lemon:writtenRep "discover"@en ];
 lemon:synBehavior   [ a lexinfo:TransitiveFrame;
                       lexinfo:subject       :arg1;
                       lexinfo:directObject  :arg2 ];
 lemon:sense         [ lemon:reference dbpedia:discoverer;
                       lemon:subjOfProp      :arg2;
                       lemon:objOfProp       :arg1 ].
```



Pattern based relation extraction, learning patterns (2)



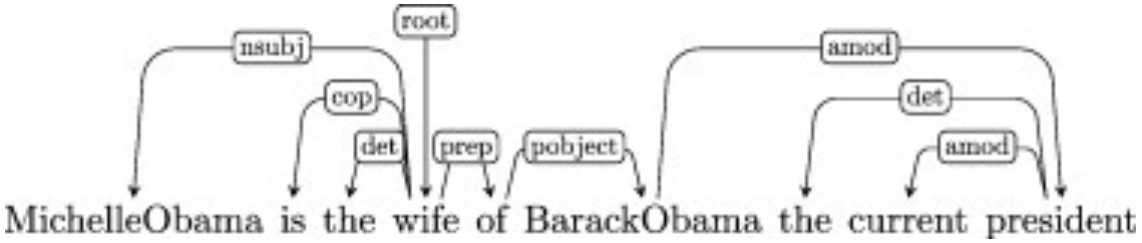
Dbpedia:spouse

```
(res:Barack_Obama, dbpedia:spouse, res:Michelle_Obama)
(res:Hilda_Gadea,  dbpedia:spouse, res:Che_Guevara)
(res:Mel_Ferrer,   dbpedia:spouse, res:Audrey_Hepburn)
```

Find all lexicalizations of the entities: Michelle Obama, Mrs. Obama, Michelle Robinson …

Michelle Obama is the wife of *Barack Obama*, the current president.

```
[ (Michelle,NNP), (Obama,NNP),
  (is,VBZ), (the,DT), (wife,NN), (of,TO),
  (Barack,NNP), (Obama,NNP),
  (the,DT), (current,JJ), (president,NN) ]
```

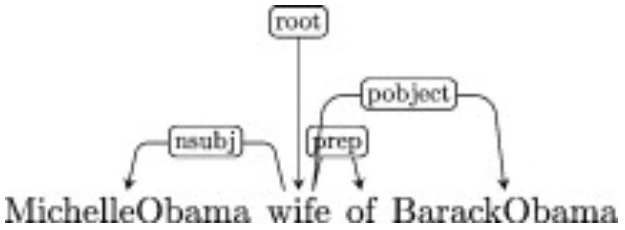


Pattern based relation extraction, learning patterns (3)

- Pattern = shortest path btw the 2 entities in the dependency graph

[MichelleObama (subject), wife (root), of (preposition), BarackObama (object)]

- Lexical entry in the ontology




```
:wife a lemon:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  lemon:canonicalForm [ lemon:writtenRep "wife"@en ] ;
  lemon:synBehavior   [ rdf:type lexinfo:NounPPFrame ;
                        lexinfo:copulativeArg       :x_subj ;
                        lexinfo:prepositionalObject :y_pobj ] ;
  lemon:sense         [ lemon:reference dbpedia:spouse;
                        lemon:subjOfProp :x_subj ;
                        lemon:objOfProp  :y_pobj ] .

:y_pobj lemon:marker [ lemon:canonicalForm
                       [ lemon:writtenRep "of"@en ] ] .
```

Extraction de relations

- Pour apprendre des patrons
  - Exploiter les graphes de dépendance
  - Exploiter des ressources pour amorcer avec des couples en relation
  - Dépendance au domaine
- Apprendre des régularités
  - Corpus de grande taille à très grande taille
  - liste ouverte de relations, d'entités et de classes relations
  - Disposer de corpus annotés -> supervised learning
  - Très grands corpus, étude des voisinages de mots et de leur similarités -> unsupervised learning