

Ahmed Halioui  
Laboratoire de bioinformatique  
UQÀM

# EXTRACTION DE L'INFORMATION PHYLOGÉNÉTIQUE

Mars 2014

# PLAN

- **Cadre du travail**
- **État de l'Art**
  - Extraction des entités nommées
  - Extraction des relations
- **Implémentation**
- **Conclusion et perspectives**

# 1. CADRE DU TRAVAIL

Système de  
recommandation  
phylogénétique

# 1. INTRODUCTION

- Extraction de connaissances phylogénétiques à partir des textes pour enrichir une ontologie de base PhylOnt8
  - Concepts = termes
    - *Neighbor joining, DAMBE, HKY*
  - Propriétés = relations
    - Is-a, is-used-in, is-parameter
- Ontologie
  - Phylont v8 (10-02-2012)
    - Bioportal
    - Métriques :
      - MappingsOF CLASSES: 147
      - NUMBER OF INDIVIDUALS: 2
      - NUMBER OF PROPERTIES: 38
      - MAXIMUM DEPTH: 6

## Information Extraction

### Named Entity Recognition

The task of  
classifying tokens  
in text documents

### Relation Extraction

The extraction of  
relations between  
formerly extracted  
named entities

# 1. SOLUTION PROPOSÉE

1. Extraction du corpus
2. Préparation de données
  - a. Segmentation du corpus
  - b. Catégorisation de termes
3. Extraction des entités nommées
  - a. Extraction des attributs-contextes
  - b. Génération des motifs d'extraction
  - c. Application d'un modèle probabiliste d'annotation
4. Extraction de relations
  - a. Extraction des attributs-contextes
  - b. Génération des motifs d'extraction
  - c. Application d'un modèle probabiliste d'extraction de relations

# 1. OBJECTIF

- Extraction des entités nommées
- Extractions de relations entre les entités
- Découvrir des événements plus larges



Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
DNA fragments	is_used_in	DNADIST
DNADIST	studied	Chagas disease
...	...	...

Information non structurée :  
Résumé (humain)

Information structurée :  
Résumé (machine)

## 2. EXTRACTION DES ENTITÉS NOMMÉES (NER)

### Information Extraction

#### Named Entity Recognition

The task of  
classifying tokens  
in text documents

#### Relation Extraction

The extraction of  
relations between  
formerly extracted  
named entities

État de  
l'Art

## 2. RECONNAISSANCE DES ENTITÉS NOMMÉES - DÉFINITION

### ■ NER ou NERC

- Sous-tâche d'un système d'extraction d'information
- Apprendre à reconnaître et classifier des éléments dans des catégories prédéfinies
  - person, location and organisation, ...
  - taxonomy, proteins, ...
- Métonymie, structure, formatage, ...

### ■ État de l'Art

\* Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigations*, 30(1), 3-26. [Cité 532 fois](#)

### ■ Vers le traitement de l'information multimédias



## 2. CHALLENGES SCIENTIFIQUES

- HUB-4 (N. Chinchor et al. 1998)
- MUC-7 and MET-2 (N. Chinchor 1999),
- IREX (S. Sekine & Isahara 2000),
- CONLL (E. Tjong Kim Sang 2002, E. Tjong Kim Sang & De Meulder 2003),
- ACE (G. Doddington et al. 2004)
- HAREM (D. Santos et al. 2006)
- CLEFFER
  - Patent documents
  - Medline abstract titles
  - EMEA corpus

## 2. RECONNAISSANCE DES ENTITÉS NOMMÉES - PRINCIPE

- Tâche :
  - 1. Trouver des entités
  - 2. Classifier par type
- Exemple (abstract\_217.txt)

The systematics of **lobose testate amoebae (Arcellinida)**, a diverse group of shelled free-living **unicellular eukaryotes**, is still mostly based on morphological criteria such as shell shape and composition. Few molecular **phylogenetic studies** have been performed on these organisms to date, and their **phylogeny** suffers from typical under-sampling artefacts, resulting in a still mostly unresolved **tree**. In order to clarify the **phylogenetic relationships** among **arcellinid testate amoebae** at the inter-generic and inter-specific level, and to evaluate the validity of the criteria used for taxonomy, we amplified and sequenced the **SSU rRNA gene** of nine taxa - **Diffflugia bacilliarum**, **D. hiraethogii**, **D. acuminata**, **D. lanceolata**, **D. achlora**, **Bullinularia gracilis**, **Netzelia oviformis**, **Physochila griseola** and **Cryptodifflugia oviformis**. Our results, combined with existing data demonstrate the following: 1) Most **arcellinids** are divided into two major clades, 2) the genus **Diffflugia** is not monophyletic, and the genera **Netzelia** and **Arcella** are closely related, and 3) **Cryptodifflugia** branches at the base of the **Arcellinida** clade. These results contradict the traditional taxonomy based on shell composition, and emphasize the importance of general shell shape in the taxonomy of **arcellinid testate amoebae**.

# 3.1. APPRENTISSAGE

## ■ Approche classique

- Règles Handcrafted (écrits à la main)
- S. Sekine and Nobata (2004)

## ■ Apprentissage supervisé

- “Tagging words of a test corpus when they are annotated as entities in the training corpus.”\*
- Attributs discriminant des exemples positifs et négatifs
- Approches
  - Hidden Markov Models (HMM) (D. Bikel et al. 1997),
  - Maximum Entropy Models (ME) (A. Borthwick 1998),
  - Support Vector Machines (SVM) (M. Asahara & Matsumoto 2003),
  - **Conditional Random Fields (CRF) (A. McCallum & Li 2003).**
- Évaluation
  - Rappel
    - vocabulary transfer (indicateur) : proportion de mots, sans répétitions, apparant dans le corpus d’entraînement et de test
  - Précision

\* Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigations*, 30(1), 3-26. [Cité 532 fois](#)

# 3.1. APPRENTISSAGE

## ■ Apprentissage semi-supervisé

### ■ Découvrir des nouveaux contextes

### ■ Approches

- **S. Brin (1998)** : attributs lexicaux implémentés dans des expressions régulières
- **M. Collins and Singer (1999)** : motifs {spelling, context}
  - "If the spelling contains "Mr." then it is a Person;"
- **M. Collins and Singer and R. Yangarber et al. (2002)** : motifs et règles pour les exemples négatifs (One type Against All)
- **E. Riloff and Jones (1999)** : Bootstrapping ; accumuler des motifs à partir d'un ensemble d'apprentissage.
- **A. Cucchiarelli and Velardi (2001)** : relations syntactiques.
- **M. Pasca et al. (2006)** : similarité des distributions
  - Par exemple, pour un motif : "X est né en Novembre" et une base de synonymes {Mars, Octobre, Avril, ...}, des nouveaux motifs peuvent être générés comme "X est née en Mars ».

\* Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigations*, 30(1), 3-26.

# 3.1. APPRENTISSAGE

## ■ Apprentissage non-supervisé

### ■ Similarité du contexte

- Ressources lexicales (WordNet, Wikipedia, etc.)
- Motifs lexicaux
- Statistiques (TFIDF)

### ■ Approches

- **E. Alfonseca and Manandhar (2002)** : Assigner un “topic” à chaque Synset à partir des co-occurrences fréquents des mots dans un corpus.
- **R. Evans (2003)** : identification des hyponymes/hypernymes à partir des séquences des mots écrits en Majuscules et autres motifs.

\* Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigations*, 30(1), 3-26.

## 3.2. ATTRIBUTS

- **Attributs de niveau Mot**
  - Casse
  - Ponctuation
  - Chiffres
  - Caractères
  - Morphologie (stem, préfix, suffixe, etc.)
  - Parties du discours (nom, verbe adjectif, etc.)
  - Autres motifs (longueur, motifs généralisés, etc.)
- **Attributs sur les listes des entités nommées (Gazetteer, lexicon, dictionnaire)**
  - Listes : stop-words, abréviations, synonymes, etc.
  - Liste des entités : classe de types, type, sous-types
- **Attributs des documents**
  - Occurrences multiples : Uppercases/lowercases, Anaphore, coréférence, etc.
  - Syntaxe locale : phrase, paragraphe, etc.
  - Mera-information : XML, List, figure, table, etc.
  - Fréquences : mots, cooccurrences, etc.

## 3.3. ÉVALUATION

### ■ Types d'erreurs\*

\* Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigations*, 30(1), 3-26.

**Table 4: NERC type of errors.**

Correct solution	System output	Error
Unlike	<ENAMEX TYPE="LOCATION"> Unlike </ENAMEX>	The system hypothesized an entity where there is none.
<ENAMEX TYPE="PERSON"> Robert </ENAMEX>	Robert	An entity was completely missed by the system.
<ENAMEX TYPE="PERSON"> John Briggs Jr </ENAMEX>	<ENAMEX TYPE="ORGANIZATION"> John Briggs Jr </ENAMEX>	The system noticed an entity but gave it the wrong label.
<ENAMEX TYPE="ORGANIZATION"> Wonderful Stockbrokers Inc </ENAMEX>	<ENAMEX TYPE="ORGANIZATION"> Stockbrokers </ENAMEX>	A system noticed there is an entity but got its boundaries wrong.
<ENAMEX TYPE="LOCATION"> New York </ENAMEX>	<ENAMEX TYPE="PERSON"> in New York </ENAMEX>	The system gave the wrong label to the entity and got its boundary wrong.

## 3.3. ÉVALUATION

### ■ Les scores MUC

- TYPE : trouver le bon type ;
- TEXT : trouver le bon texte ;
- COR : nombre des réponses correctes (vrais positifs) ;
- ACT : nombre des annotations trouvées (vrais positifs + faux positifs) ;
- POS : nombre des entités possible dans une solution (vrais positifs + faux négatifs);
- Précision :  $COR / ACT$
- Rappel :  $COR / POS$
- MAF : f-mesure



# 3.4 APPLICATIONS – STANFORD NER

## ■ Stanford NER

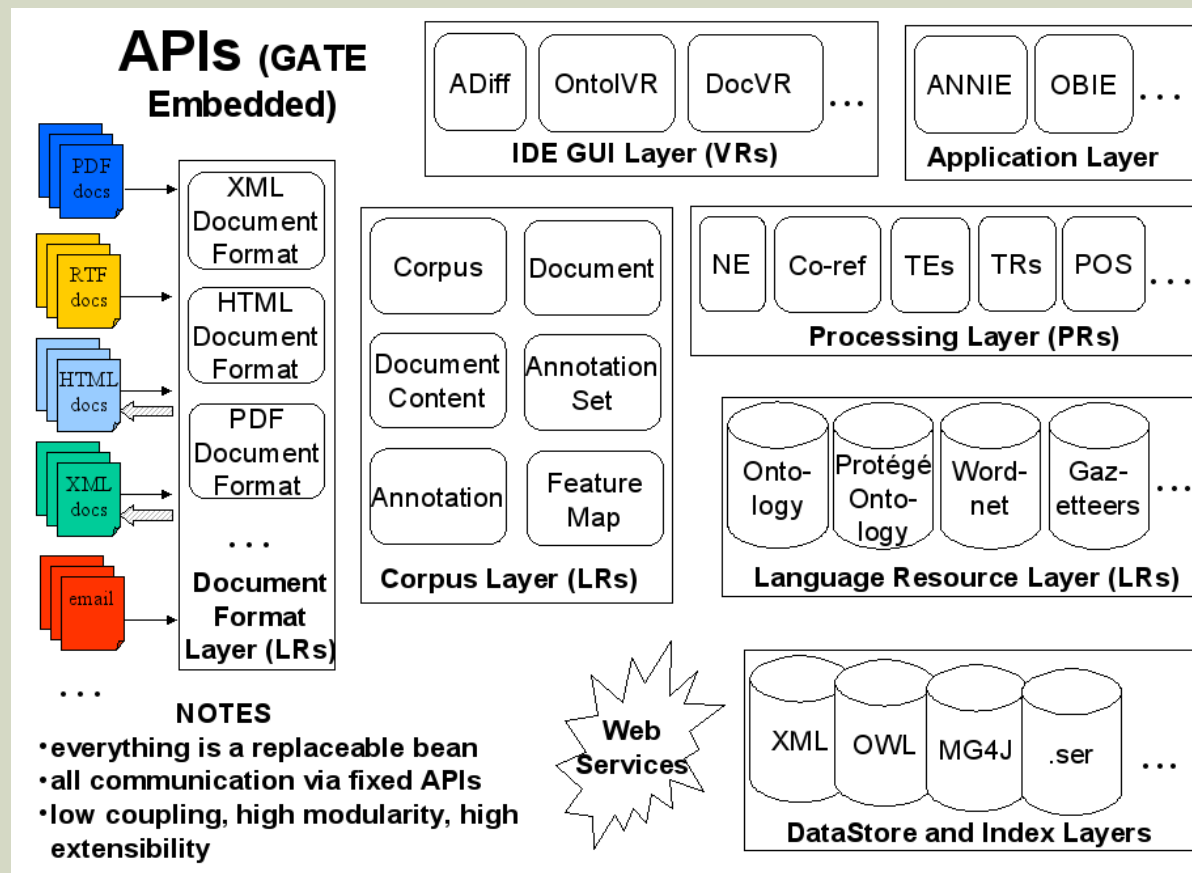
### ■ Basé sur un CRF

- <http://nlp.stanford.edu/software/CRF-NER.shtml>

Stanford NLP Named Entity Recognition Results







Corpus		# Word Tokens		# Entities		# Features		Exact Match Score (conlleval)			Technology		Notes
Name	Language	Train	Test	Types	Instances	$\Phi(X)$	$\lambda/f(X,Y)$	Prec	Rec	$F_1$	Classifier	Properties file/flag	
CoNLL 2002	Dutch news testa (devset)	218737	37761	4	2616	838524	4192620	78.99%	77.33%	78.15%	pure CMM	-goodCoNLL	1, 3, 5, 7
CoNLL 2002	Dutch news testb	218737	68994	4	3941	838559	4192795	80.48%	78.96%	79.71%	pure CMM	-goodCoNLL	1, 3, 5, 7
CoNLL 2002	Spanish news testa (devset)	273037	52923	4	4352	776511	3882555	78.01%	76.19%	77.09%	pure CMM	-goodCoNLL	1, 3, 5, 7
CoNLL 2002	Spanish news testb	273037	51533	4	3559	776444	3882220	81.24%	81.03%	81.14%	pure CMM	-goodCoNLL	1, 3, 5, 7
CoNLL 2003	English news testa (devset)	219553	51578	4	5942	738378	3691890	91.37%	91.22%	91.29%	pure CMM	-goodCoNLL	1, 5, b
CoNLL 2003	English news testa (devset)	219554	51578	4	5942			92.15%	92.39%	92.27%	postprocessed CMM		1, 2, 4
CoNLL 2003	English news testb	219553	46666	4	5648	738378	3691890	85.65%	85.41%	85.53%	pure CMM	-goodCoNLL	1, 5, b
CoNLL 2003	English news testb	219554	46666	4	5648			86.12%	86.49%	86.31%	postprocessed CMM		1, 2, 4
CoNLL 2003	German news testa (devset)	220189	51645	4	4833	1079044	5395220	77.12%	61.37%	68.35%	pure CMM	-goodCoNLL	1, 3, 5, 6, 7, a
CoNLL 2003	German news testa (devset)	220189	51645	4	4833			75.36%	60.36%	67.03%	postprocessed CMM		1, 2, 3, 4
CoNLL 2003	German news testb	220189	52098	4	3673	1079037	5395185	79.23%	63.65%	70.59%	pure CMM	-goodCoNLL	1, 3, 5, 6, 7, a
CoNLL 2003	German news testb	220189	52098	4	3673			80.38%	65.04%	71.90%	postprocessed CMM		1, 2, 3, 4
CoNLL 2003	English news testa (devset)	219553	51578	4	5942	616918	11532202	91.64%	90.93%	91.28%	CRF (closed task)	conll.crf.chris2009.prop lob2	1, 5, c
CoNLL 2003	English news testa (devset)	219553	51578	4	5942	633786	12285708	93.28%	92.71%	92.99%	CRF (with distsim)	conll.crf.chris2009.prop lob2 distsim	1, 5, c
CoNLL 2003	English news testb	219553	46666	4	5648	633786	12285708	88.21%	87.68%	87.94%	CRF (with distsim)	conll.crf.chris2009.prop lob2 distsim	1, 5, c

## 3.4. APPLICATIONS – GATE IE



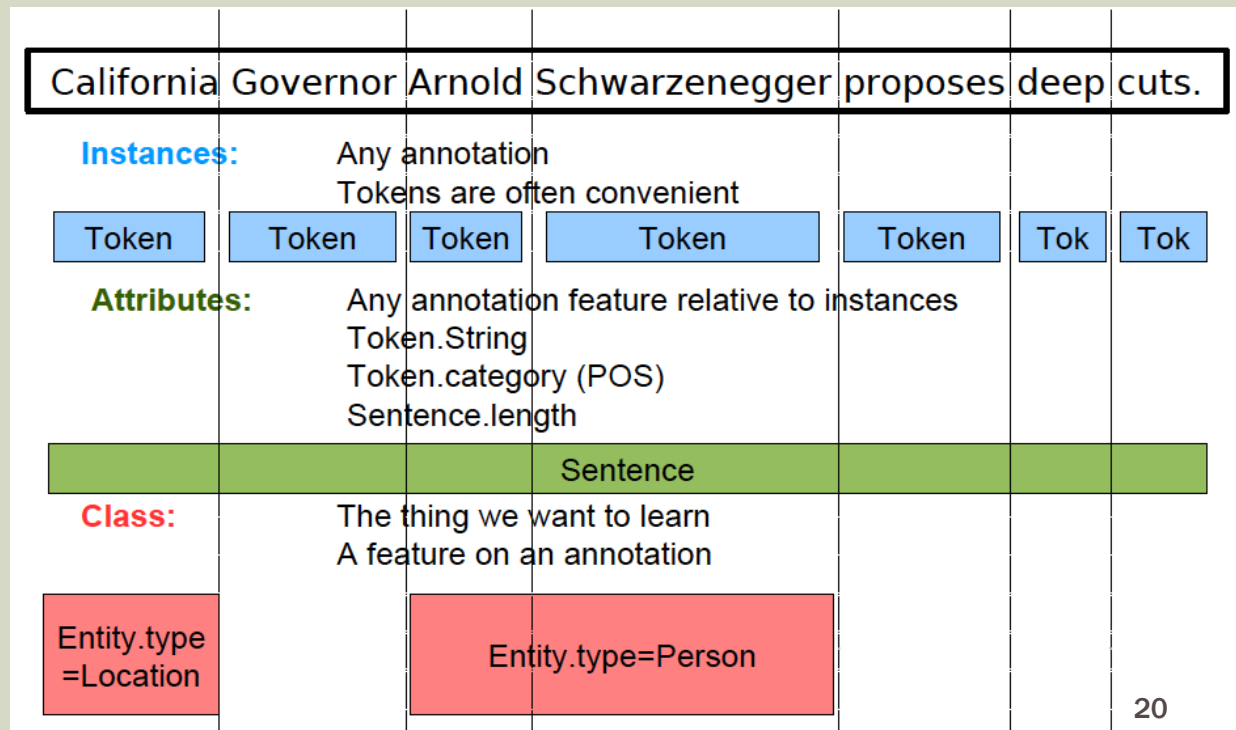
## 3.4. APPLICATIONS – GATE ANNIE

- ANNIE (Nearly-New Information Extraction system)

Selected Processing resources		
!	Name	Type
	ANNIE English Tokeniser_000D6	ANNIE English Tokeniser
	ANNIE Gazetteer_000D9	ANNIE Gazetteer
	ANNIE Sentence Splitter_000DA	ANNIE Sentence Splitter
	ANNIE POS Tagger_000DD	ANNIE POS Tagger
	ANNIE NE Transducer_000DE	ANNIE NE Transducer
	ANNIE OrthoMatcher_000DF	ANNIE OrthoMatcher

## 3.4. APPLICATIONS – GATE LEARNING

- Chunk recognition
- Text classification
- Relation annotation



## 3.4. APPLICATIONS – GATE IE

### ■ Projets Gate

- **Parallel IE**, Merck KGaA, Darmstadt, Germany - Information Extraction on a Linux cluster **for bio-medical text mining and indexing**.
- **Medical Informatics**, University of Pittsburgh, USA - Annotating surgical pathology reports using UMLS.
- **Medical Informatics**, Institute for Medical Informatics and Biometry, University of Rostock, Germany - **Analyzing MEDLINE abstracts to extract causal functional relations, which are essential for the construction of genetic networks, as a step towards characterization of diseases**.
- **BioRat**, University College, London, U.K. (Corney, 2004) - A general-purpose information extraction tool designed to be used by biologists to data-mine text from journals. It has been successfully applied to **protein-protein interaction discovery** and more projects are underway in several other areas. It uses GATE at its core, while also providing tools to design new templates, edit gazetteers and to download full-length papers from the web. The software is available for academic use, and is part of an ongoing research project.
- **InESBi**, Institute for Medical Informatics and Biometry, University of Rostock, Germany - the information extraction for this **structural biology project is aimed at the 'material and method' part of the structural biology publications**. The purpose of this project is to populate a database. Some of the pieces of information for the database are retrieved from structured files named PDB. The material and method used for experiments are not in PDB files. Thus, the intent is to extract that information from the text of the publications.

## 3.5. CONCLUSION

	Principe	Auteurs
Approche classique	Règles écrits à la main	S. Sekine and Nobata (2004)
Apprentissage supervisé	Hidden Markov Models (HMM)	(D. Bikel et al. 1997)
	Maximum Entropy Models (ME)	(A. Borthwick 1998)
	Support Vector Machines (SVM)	(M. Asahara & Matsumoto 2003)
	Conditional Random Fields (CRF)	(A. McCallum & Li 2003)
Apprentissage non supervisé	Assigner un “topic” à chaque Synset à partir des co-occurrences fréquents des mots dans un corpus.	E. Alfonseca and Manandhar (2002)
	identification des hyponymes/hyponymes à partir des séquences des mots écrits en Majuscules et autres motifs.	R. Evans (2003)
Apprentissage semi-supervisé	attributs lexicaux implémentés dans des expressions régulières	S. Brin (1998)
	motifs {spelling, context}	M. Collins and Singer (1999)
	motifs et règles pour les exemples négatifs	M. Collins and Singer and R. Yangarber et al. (2002)
	Bootstrapping	E. Riloff and Jones (1999) :
	relation syntactiques.	A. Cucchiarelli and Velardi (2001)
	similarité des distributions	M. Pasca et al. (2006)

\* Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigations*, 30(1), 3-26. [Cité 532 fois](#)

# 4. EXTRACTION DES RELATIONS

## Information Extraction

### Named Entity Recognition

The task of classifying tokens in text documents

### Relation Extraction

The extraction of relations between formerly extracted named entities

État de l'Art

# 4.1. EXTRACTION DE RELATIONS

## ■ Exemple (abstract\_217.txt)

The systematics of **lobose testate amoebae (Arcellinida)**, a diverse group of shelled free-living **unicellular eukaryotes**, is still mostly based on morphological criteria such as shell shape and composition. Few molecular **phylogenetic studies** have been performed on these organisms to date, and their **phylogeny** suffers from typical under-sampling artefacts, **resulting** in a still mostly unresolved **tree**. In order to **clarify** the **phylogenetic relationships among arcellinid testate amoebae** at the inter-generic and inter-specific level, and to evaluate the validity of the criteria used for taxonomy, we **amplified** and **sequenced** the **SSU rRNA gene** of nine taxa - **Diffugia bacilliarum**, **D. hiraethogii**, **D. acuminata**, **D. lanceolata**, **D. achlora**, **Bullinularia gracilis**, **Netzelia oviformis**, **Physochila griseola** and **Cryptodiffugia oviformis**. Our results, combined with existing data demonstrate the following: 1) Most **arcellinids** are divided into two major clades, 2) the genus **Diffugia** is not monophyletic, and the genera **Netzelia** and **Arcella** are closely related, and 3) **Cryptodiffugia** branches at the base of the **Arcellinida** clade. These results contradict the traditional taxonomy based on shell composition, and emphasize the importance of general shell shape in the taxonomy of **arcellinid testate amoebae**.

Taxonomy

Method

Data\_type

Relations\_candidates



## 4.2. TYPES DE RELATIONS

- Bases de relations
  - ACE 2003
    - Role, At, Social, ...
  - Freebase
    - People, film, book, ...
  - Gene Ontology
    - Is-a, part=of, regulates
- NYU'S Proteus
  - Disease outbreaks from The New York Times
- WordNet
  - Incomplet ☹

## 4.4. MOTIFS ÉCRITS À LA MAIN

### ■ Motifs *Handcrafted*

- *NYU Proteus system (1997)*
- *Intuition from Hearst (1992)*
  - *Hyponymes*
    - *Such as, or other, etc.*
  - *...*

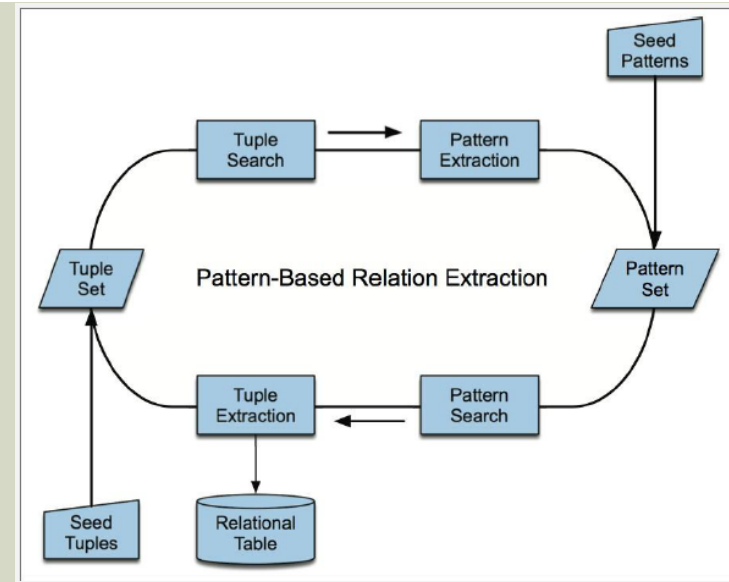
- *(-) Conception des motifs écrits à la main*
- *(-) Concevoir tous les types de relations*
- *(-) Accuracy -66 %*

```
;;; For <company> appoints <person> <position>

(defun pattern appoint
  "np-sem(C-company)? rn? sa? vg(C-appoint) np-sem(C-person) ', '?'
  to-be? np(C-position) to-succeed?:
  company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attributes
  position-at=8.attributes |
  ..."
  (defun when-appoint (phrase-type)
    (let ((person-at (binding 'person-at))
          (company-entity (entity-bound 'company-at))
          (person-entity (essential-entity-bound 'person-at 'C-person))
          (position-entity (entity-bound 'position-at))
          (predecessor-entity (entity-bound 'predecessor-at))
          new-event)
      (not-an-antecedent position-entity)
      ;; if no company is specified for position, use agent
      ...
    )
  )
)
```

## 4.5. BOOTSTRAPPING

- Si les textes sont "peu" annotés mais
  - Échantillons des instances de relations / motifs de relations
  - Beaucoup de textes non annotés (Web)
- Méthode semi-supervisée
- Exemple.
  - Relation cible : is\_result
  - Tuple (échantillon) : (phylogeny, tree)
  - Chercher sur le Web « phylogeny » et « tree »
    - "The Tree of Life then **represents** the phylogeny."
    - »A phylogeny is used to **construct** the tree of ..."
  - Utiliser ces motifs pour chercher des nouveaux tuples



## 4.6. BOOTSTRAPPING

### ■ Applications

- DIPRE (Brin, 1998)
- Snowball (Agichtein & Gravano, 2000)

### ■ Inconvénients

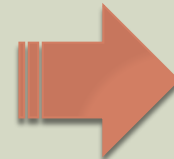
- (-) Sensibilité à l'échantillon de départ des échantillons de relations/motifs
- (-) Un grand problème de changement de contexte/sémantique dans chaque itération
- (-) Précision faible
- (-) Beaucoup de paramètres
- (-) Pas d'interprétation probabiliste

## 4.7. APPROCHE SUPERVISÉE

- Définition préalable des classes de relations
- Définition d'un ensemble d'apprentissage
- Définition des attributs

Features commonly used in relation classification:

- Lightweight features — require little pre-processing
  - Bags of words & bigrams between, before, and after the entities
  - Stemmed versions of the same
  - The types of the entities
  - The distance (number of words) between the entities
- Medium-weight features — require base phrase chunking
  - Base-phrase chunk paths
  - Bags of chunk heads
- Heavyweight features — require full syntactic parsing
  - Dependency-tree paths
  - Constituent-tree paths
  - Tree distance between the entities
  - Presence of particular constructions in a constituent structure



- Choisir un modèle de classification : SVM, MaxEnt, NB, ....
- Évaluation des résultats

## 4.7. APPROCHE SUPERVISÉE - ATTRIBUTS

- Mots
  - Sac-de-mots
  - Mots-entre-entités
  - Mots-avant-après
  - (+) bonne précision mais (-) faible rappel
- Types des entités nommées
  - Entreprise, Lieu, Personne, etc.
  - (+) aide à améliorer le taux de rappel
- Niveau de discours
  - POS : nom, adverbe, verbe, ....
  - (+) effet non considérable des fois

## 4.7. APPROCHE SUPERVISÉE - ATTRIBUTS

## ■ Chevauchement

- Nombre de mots entre les entités nommées
- Attributs conjoints
- (+) aide à améliorer le rappel mais (-) affaiblit la précision

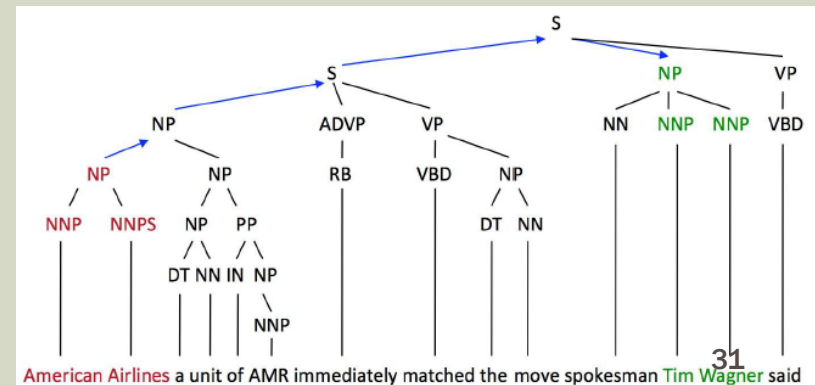
## ■ Phrase chunking

- Segmenter une phrase en morceaux

- Entêtes avant et après [NP American Airlines], [NP a unit] [PP of] [NP AMR], [ADVP immediately]
- Chemin [VP matched] [NP the move], [NP spokesman Tim Wagner] [VP said].
- (+) aide à améliorer le rappel et la précision

## ■ Attributs syntactiques

- Arbres de dépendances
- (+/-) impact négligeable



## 4.7. APPROCHE SUPERVISÉE - CLASSIFIEURS

- Classifieurs multi-classes (un contre un, une contre les autres):
  - Zhou et al. 2005 : SVM
  - MaxENT (régression logistique multi-classes)
  - Naives Bayes
  - Etc.
- Classifieur Multi-instances multi-classes
  - Exemples
    - "Balzac" et "France" peuvent exprimer les relations NéEn ou MortEn.
    - "Tree" et "ClustalW2" peuvent exprimer les relations "EstEntrée" ou "EstSortie", elles peuvent exprimer aussi "estAlignéPar".

Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. (2012, July). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 455-465). Association for Computational Linguistics.



# 4.8. SUPERVISION DISTANTE

Zajac, M., & Przepiórkowski, A. (2013, January). Distant supervision learning of DBPedia relations. In *Text, Speech, and Dialogue* (pp. 193-200). Springer Berlin Heidelberg. [Cité 5 fois](#)

## ■ Hypothèse

- Si deux entités appartiennent à une relation, alors une telle relation est décrite dans la même phrase.

## ■ Idée-clé

- Pour chaque couple d'entités :
  - 1. Chercher les phrases cibles
  - 2. Extraire des attributs
  - 3. Construire un modèle de classification

## ■ Avantages

- (+) les avantages de l'approche supervisée : information apriori, forme canonique des relations, attributs.
- (+) les avantages de l'approche non supervisée : tolérer des attributs non discriminants (noisy), pas sensible à l'ensemble d'apprentissage

## 4.8. SUPERVISION DISTANTE

### ■ Exemple. Apprentissage des hyperonymes

- 1. Extraire les phrases
- 2. Collecter les paires des entités
- 3. Chercher si la paire est de types IS\_A dans WordNet/Freebase/autres.
- 4. Parser les phrases
- 5. Extraire des motifs
- 6. Construire un modèle
- 7. Évaluer le modèle

### ■ (+) Précision élevée

### ■ (+) Les attributs syntaxiques et lexicales peuvent aider à construire des motifs pertinents

### ■ (-) Les connaissances à priori doivent être collectés et prétraités (base de relations)

# 4.8. SUPERVISION DISTANTE

## Corpus

we **amplified** and **sequenced** the **SSU rRNA gene** of nine taxa - *Diffugia bacilliarum*, *D. hiraethogii*, *D. acuminata*, *D. lanceolata*, *D. achlora*, *Bullinularia gracilis*, *Netzelia oviformis*, *Physochila griseola* and *Cryptodiffugia oviformis*.

We **examined** sequence variation in one **mitochondrial** (**12S rRNA**) and three nuclear genes (**vWF**, **GHR** and **RAG1**) across all **caviomorph families**.

**Maximum Likelihood**, **Bayesian** and **Maximum Parsimony** analyses of a combination of over two and a half kilobases of nuclear (PDC, Rag-1) and **mitochondrial** (ND2, ND4, tRNA) sequence data all **identified** four distinctive lineages within *Oedura s.l.*

## Source

is used in: (**SSU rRNA gene**, *Diffugia bacilliarum*)  
is used in: (**SSU rRNA gene**, *D. hiraethogii*)  
is used in: (**SSU rRNA gene**, *D. acuminata*)  
is used in: (**SSU rRNA gene**, *Diffugia bacilliarum*)

is used in: (**mitochondrial**, *caviomorph families*)

is used in: (**mitochondrial**, *Oedura s.l.*)

## Ensemble d'apprentissage

is used in: (**SSU rRNA gene**, *Diffugia bacilliarum*)

Label: is\_used\_in

Feature : **amplified and sequenced** the X of nine taxa - Y

is used in: (**SSU rRNA gene**, *Diffugia bacilliarum*)

Label: is\_used\_in

Feature : We **amplified and sequenced** the X of nine taxa - *Diffugia bacilliarum*, Y

is used in: (**mitochondrial**, *caviomorph families*)

Label: is\_used\_in

Feature : We **examined** sequence variation of X (12S rRNA) and three nuclear genes (vWF, GHR and RAG1) across all Y.

- L'ensemble négatif est construit en construisons des relations invalides à partir les différentes combinaisons possibles entre les entités nommées.

## 4.9. APPRENTISSAGE NON SUPERVISÉ

### ■ KnowItAll (Etzioni et al, 2005)

#### ■ Entrée :

- Une base des prédicats et classes cibles
  - Uses(Data\_type, Taxon)
- Utiliser les motifs Hearst pour trouver toutes instances de classes
  - Such as, an other ,as, ....
- Utiliser des motifs pour trouver des relations
  - "We **RELATION** Data\_type across Taxon"
  - "Data\_type **RELATION** Taxon. »
- Ajouter la relation apprise dans la base des prédicats

## 4.9. APPRENTISSAGE NON SUPERVISÉ

### ■ TextRunner (Banko et al,2007)

#### ■ 1. Self-supervised learner

- Parser des phrases (POS)
- B. Extraire tous les tuples
- C. Étiqueter chaque tuple en se basant sur les attributs POS
  - Exemple :
    - Positif is la dépendance est courte (<3) et les NP (noms propres) ne sont pas des pronoms
- D. Créer un modèle Naïf Bayes sur les tuples créés
  - En utilisant des attributs lightweight : POS, Stopwords, ...

#### ■ 2. Single-pass extractor

- A. Utiliser un POS tagger
- B. Utiliser un Base Noun Phrase chunker
- C. Extraire toutes les chaines de caractères entre les NP
- D. Utiliser des règles heuristiques pour simplifier les chaines de caractères
- E. Passer les tuples créés sur un classifieur Naïf Bayes.

#### ■ 3. Redundancy-based assessor

- A. Compter les toutes instances d'un tuple
- B. Calculer la vraisemblance de chaque tuple

## 4.9. APPRENTISSAGE NON SUPERVISÉ

- **DIRT (Lin & Pantel 2003):** Discovery of Onference Rules from Text
  - En utilisant les chemins de dépendances de MINIPAR entre les paires de noms, il s'agit de retrouver les chemins similaires
  - Si deux chemins représentent des contextes similaires, alors les significations tendent d'être similaires. (**FAUX !**)
  - Exemple : les top 20 chemins similaires à "X solves Y"

Y is solved by X	Y is resolved in X
X resolves Y	Y is solved through X
X finds a solution to Y	X rectifies Y
X tries to solve Y	X copes with Y
X deals with Y	X overcomes Y
Y is resolved by X	X eases Y
X addresses Y	X tackles Y
X seeks a solution to Y	X alleviates Y
X do something about Y	X corrects Y
X solution to Y	X is a solution to Y

- **(-)** ambiguïtés de chemins

## 4.9. APPRENTISSAGE NON SUPERVISÉ

- Une amélioration de DIRT (Yao et al. 2012)
  - Extraire et filtrer les tuples (entité, chemin, entité)
  - Construire les représentations/attributs pour chaque tuple
    - Sac-de-mots, mot, catégories de documents, catégories de phrase (LDA topic model)
  - Regrouper les tuples pour chaque chemin dans des clusters
    - Appliquer le modèle LDA topic (regrouper par topic)
  - Chercher les relations sémantiques dans chaque cluster
    - Clustering hiérarchique agglomératif
      - Similarité entre les vecteurs attributs

# 4.10. CONCLUSION

	Principe	Auteurs
Approche classique	Règles écrits à la main	NYU Proteus system (1997) Hearst (1992)
Apprentissage semi-supervisé (Bootstrapping)	Chercher sur le Web les tuples de relations	DIPRE (Brin, 1998) Snowball (Agichtein & Gravano, 2000)
Apprentissage supervisé	SVM	(Zhou et al., 2005)
	MaxENT	
	Naive Bayes	
	Multi-Instances muli-classes	(Surdeanu et al., 2012)
Supervision distante	Le contexte d'une relation est définie dans une phrase.	(Zajac, M., & Przepiórkowski, A., 2013)
Apprentissage non supervisé	<b>KnowItAll</b> : utiliser les motifs de Hearst pour chercher les relation dans une phrase	(Etzioni et al, 2005)
	<b>TextRunner</b> : il s'agit d'identifier les tuples, créer des motifs pour les chemins entre les entités nommées et éliminer les tuples redondants	(Banko et al,2007)
	DIRT : Si deux chemins représentent des contextes similaires, alors les significations tendent d'être similaires.	(Lin & Pantel 2003)
	Une amélioration de DIRT	(Yao et al. 2012)



# 5. IMPLÉMENTATION

Check  
Point.

# 5.1 EXTRACTION DES ENTITÉS NOMMÉES

## - PRINCIPE

- Schéma général d'apprentissage
  - 1. Extraction de textes
  - 2. Classification de textes
  - 3. Création des annotations syntaxiques et lexicales
  - 4. Création des annotations du domaine phylogénétique
  - 5. Écrire des règles d'annotation génériques en utilisant les annotations syntaxique et lexicales
    - motifs d'annotation spécifique à chaque annotation phylogénétique.
  - 6. Créer un modèle d'annotation
  - 7. Évaluer le modèle crée

# 5.2 EXTRACTION DES ENTITÉS NOMMÉES

## - APPROCHE

- Approche basée sur les modules de Gate
  - 1. English Tokenizer
    - Adapté pour des tokens en anglais
    - Ce module utilise des règles écrits en JAPE
  - 2. Sentence Splitter
    - Ce module utilise une liste des abréviations pour distinguer les points de fin de ligne.
  - 3. GENIA tagger
    - Utiliser des POS tagging avec un réseau de dépendance cyclique et un classifieur probabiliste local
  - 4. Gate Morphological Analyzer
    - Stemming : ajouter de la flexibilité au gazetteer
  - 5. Phylogenetic Gazetteer
    - 12 classes: Aa, Df, Dt, Dis, Gen, Mth, Mdl, Nuc, Prog, Prot, Src, Tax
  - 6. ANNIE Gazetter
    - 48 classes/types/sous-types d'annotation
  - 7. Phylogentic Transducer (Named Entities)
    - Établir des règles spécifiques pour chaque classe d'entité nommée
  - 8. Batch PR Learning (Named Entities)
    - Choisir un modèle d'apprentissage et le bien configurer

# 5.3 EXTRACTION DES ENTITÉS NOMMÉES - ANNOTATIONS

## ■ Entités nommées touchant l'analyse phylogénétique

- GENIA (Proteins, ADN/ARN, Cellules)
- Maladies (KEGG diseases)
- Annotations d'analyse phylogénétiques
  - Formats de données,
  - Types de données,
  - Nom des Sources de données,
  - Méthodes phylogénétiques,
  - Modèles d'évolution,
  - Noms des programmes,
  - Noms des packages phylogénétiques,
  - Nom des espèces.

## ■ Autres (voir

<http://gate.ac.uk/sale/tao/splitch16.html#sec:domain-creole:biomed:genia>)

- ABNER; MetaMap, Gspell, BADRex, MiniChem§Drig tagger, AbGene, PennBioTagger, MutationFinder, NormaGene

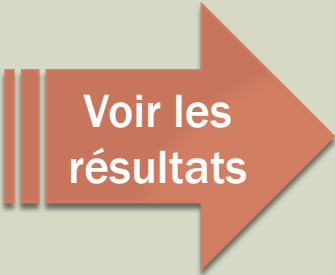
GENIA On NLPA dataset

Entity Type	Recall	Precision	F-score
Protein	81.41	65.82	72.79
DNA	66.76	65.64	66.20
RNA	68.64	60.45	64.29
Cell Line	59.60	56.12	57.81
Cell Type	70.54	78.51	74.31
Overall	75.78	67.45	71.37

Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. I. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in informatics* (pp. 382-392). Springer Berlin Heidelberg. [Cité 301 fois](#)

# 5.4 EXTRACTION DES ENTITÉS NOMMÉES - RÈGLES

- Il s'agit de spécifier des motifs et règles de reconnaissances de chaque annotations en utilisant les annotations lexicales et syntaxique
  - Nom des maladies
    - Utiliser les règles de Drug tagger
  - Nom des gènes
    - Utiliser les règles de GENIA
  - **Formats de données**
    - ...
  - **Types de données**
    - ...
  - Nom des Sources de données
    - ...
  - **Méthodes phylogénétiques**
    - ...
  - **Modèles d'évolution**
    - ...
  - **Noms des programmes**
    - ...
  - **Noms des packages phylogénétiques**
    - ...
  - **Nom des espèces**
    - ...
- Utiliser des motifs génériques du modèle phylogénétique
  - Les annotations phylogénétique ne peuvent pas être des noms de personnes, des adverbes, des pronoms personnels et autres POS.



Voir les  
résultats

# 5.5 EXTRACTION DES ENTITÉS NOMMÉES - MODÈLES

- CRF
  - Stanford NER
- SVM
  - GATE
- Et autres.

# 6.1 EXTRACTION DES RELATIONS - PRINCIPE

- Schéma général d'apprentissage – une seule relation à la fois
  - 1. **Idée 1\*** Création des instances de relations à partir des propriétés qui existent entre les classes d'une ontologie
    - Créer des instances d'exemples négatif en combinant les entités nommées
  - 2. Écrire des règles d'annotation génériques pour les relations
    - Par contexte, soit des phrases
    - **Idée 2\*** distinguer les relations symétrique, fonctionnelles, réflexive en appliquant des règles génériques pour les identifier.
  - 3. Créer des attributs d'apprentissage
  - 4. Créer un modèle d'apprentissage
  - 5. Évaluer le modèle d'apprentissage

# 6.1 EXTRACTION DES RELATIONS - INSTANCES

- Création des classes de relations
- Création des instances de relations
- Création des attributs pour les chemins (entre les deux entités dans la même phrase)
- Appliquer l'amélioration de DIRT en faisant un clustering par classe d'entités nommées les attributs
  - Par exemple soit le tuple (mitochondrial, caviomorph families), ils s'agit de regrouper tous les chemins de type de données et espèces dans le même groupe de verbes ...
  - Ces verbes constituent les instances de la relation ontologique
  - Il s'agit ensuite de filtrer les verbes en appliquant un clustering agglomératif

- has
  - has\_Ancessor
  - has\_Annotation
  - ▶ ■ has\_Node
  - has\_Root
  - has\_State
- has\_input
- has\_output
- has\_substitution\_model
- in\_topic
- is\_a\_approach\_of
- is\_a\_branch\_of
- is\_a\_method\_of ≡ is\_used\_in
- is\_a\_sub-class\_of
- is\_a\_technique\_of
- is\_branch\_of
- is\_classify\_as
- is\_format\_of
- is\_function\_as
- is\_identifier\_of
- is\_implemented\_by
- is\_inferred\_by
- is\_output\_of
- is\_result\_from
- is\_used\_by
- is\_used\_in ≡ is\_a\_method\_of
- is\_used\_to\_build
- uses



# 6.1 EXTRACTION DES RELATIONS - EXEMPLE

## Corpus

we **amplified** and **sequenced** the **SSU rRNA gene** of nine taxa - *Diffugia bacilliarum*, *D. hiraethogii*, *D. acuminata*, *D. lanceolata*, *D. achlora*, *Bullinularia gracilis*, *Netzelia oviformis*, *Physochila griseola* and *Cryptodiffugia oviformis*.

We **examined** sequence variation in one **mitochondrial** (**12S rRNA**) and three nuclear genes (**vWF**, **GHR**, and **RAG1**) across all **caviomorph families**.

**Maximum Likelihood**, **Bayesian** and **Maximum Parsimony** analyses of a combination of over two and a half kilobases of nuclear (PDC, Rag-1) and **mitochondrial** (ND2, ND4, tRNA) sequence data all **identified** four distinctive lineages within **Oedura s.l.**

## Instances - C1 is\_used\_in C2

(**SSU rRNA gene**, *Diffugia bacilliarum*)  
(**SSU rRNA gene**, *D. hiraethogii*)  
(**SSU rRNA gene**, *D. acuminata*)  
(**SSU rRNA gene**, *Diffugia bacilliarum*)

(**mitochondrial**, **caviomorph families**)

(**mitochondrial**, *Oedura s.l.*)

## Ensemble d'apprentissage

is used in: (**SSU rRNA gene**, *Diffugia bacilliarum*)  
Label: is\_used\_in  
Label\_instance : **amplified, sequenced**  
Feature : distance  
Feature : direction  
Feature : ...

is used in: (**SSU rRNA gene**, *Diffugia bacilliarum*)  
Label: is\_used\_in  
Label\_instance : **amplifiy, sequenced**  
Feature : distance  
Feature : direction  
Feature : ...

is used in: (**mitochondrial**, **caviomorph families**)  
Label: is\_used\_in  
Label\_instance : **examined**  
Feature : distance  
Feature : direction  
Feature : ...

# CONCLUSION



# CONCLUSION

# CE QU'IL RESTE À FAIRE

- **Résumés (2 mois)**
  - Créer des règles d'identification des entités nommées
  - Créer des règles d'identification des relations
  - Configurer les modèles d'apprentissage
  - **PUBLIER X2** :D
  - Évaluer les modèles appris
- **Matériel et méthodes (2 mois)**
  - Appliquer la même approche sur la section matériels et méthodes
  - **PUBLIER** :D
- **Et voilà, il ne reste que (6 mois)**
  - Populer l'ontologie
  - **PUBLIER** :D
  - Chercher des séquences d'analyse phylogénétique sur myExperiments
  - Créer le module de recommandation !!!
  - **PUBLIER X2** :D



# POS

CC - coordinating conjunction: ‘and’, ‘but’, ‘nor’, ‘or’, ‘yet’, plus, minus, less, times (multiplication), over (division). Also ‘for’ (because) and ‘so’ (i.e., ‘so that’).

CD - cardinal number

DT - determiner: Articles including ‘a’, ‘an’, ‘every’, ‘no’, ‘the’, ‘another’, ‘any’, ‘some’, ‘those’.

EX - existential ‘there’: Unstressed ‘there’ that triggers inversion of the inflected verb and the logical subject; ‘There was a party in progress’.

FW - foreign word

IN - preposition or subordinating conjunction

JJ - adjective: Hyphenated compounds that are used as modifiers; happy-go-lucky.

JJR - adjective - comparative: Adjectives with the comparative ending ‘-er’ and a comparative meaning. Sometimes ‘more’ and ‘less’.

JJS - adjective - superlative: Adjectives with the superlative ending ‘-est’ (and ‘worst’). Sometimes ‘most’ and ‘least’.

JJSS - -unknown-, but probably a variant of JJS

-LRB- - -unknown-

LS - list item marker: Numbers and letters used as identifiers of items in a list.

MD - modal: All verbs that don’t take an ‘-s’ ending in the third person singular present: ‘can’, ‘could’, ‘dare’, ‘may’, ‘might’, ‘must’, ‘ought’, ‘shall’, ‘should’, ‘will’, ‘would’.

NN - noun - singular or mass

NNP - proper noun - singular: All words in names usually are capitalized but titles might not be.

NNPS - proper noun - plural: All words in names usually are capitalized but titles might not be.

NNS - noun - plural

NP - proper noun - singular

NPS - proper noun - plural

PDT - predeterminer: Determiner like elements preceding an article or possessive pronoun; ‘all/PDT his marbles’, ‘quite/PDT a mess’.

POS - possessive ending: Nouns ending in ‘s’ or ‘’.

PP - personal pronoun

PRPR\$ - unknown-, but probably possessive pronoun

PRP - unknown-, but probably possessive pronoun

PRP\$ - unknown, but probably possessive pronoun, such as ‘my’, ‘your’, ‘his’, ‘his’, ‘its’, ‘one’s’, ‘our’, and ‘their’.

RB - adverb: most words ending in ‘-ly’. Also ‘quite’, ‘too’, ‘very’, ‘enough’, ‘indeed’, ‘not’, ‘n’t’, and ‘never’.

RBR - adverb - comparative: adverbs ending with ‘-er’ with a comparative meaning.

RBS - adverb - superlative

RP - particle: Mostly monosyllabic words that also double as directional adverbs.

STAART - start state marker (used internally)

SYM - symbol: technical symbols or expressions that aren’t English words.

TO - literal “to”

UH - interjection: Such as ‘my’, ‘oh’, ‘please’, ‘uh’, ‘well’, ‘yes’.

VBD - verb - past tense: includes conditional form of the verb ‘to be’; ‘If I were/VBD rich...’.

VBG - verb - gerund or present participle

VCN - verb - past participle

VBP - verb - non-3rd person singular present

VB - verb - base form: subsumes imperatives, infinitives and subjunctives.

VBZ - verb - 3rd person singular present

WDT - ‘wh’-determiner

WP\$ - possessive ‘wh’-pronoun: includes ‘whose’

WP - ‘wh’-pronoun: includes ‘what’, ‘who’, and ‘whom’.

WRB - ‘wh’-adverb: includes ‘how’, ‘where’, ‘why’. Includes ‘when’ when used in a temporal sense.