Extraction automatique de noms d'entreprises à partir de titres de presse : un exemple d'application chez ReportLinker

Marilyne Latour [1] Jocelyn Bernard [1] Corentin Regal [1]

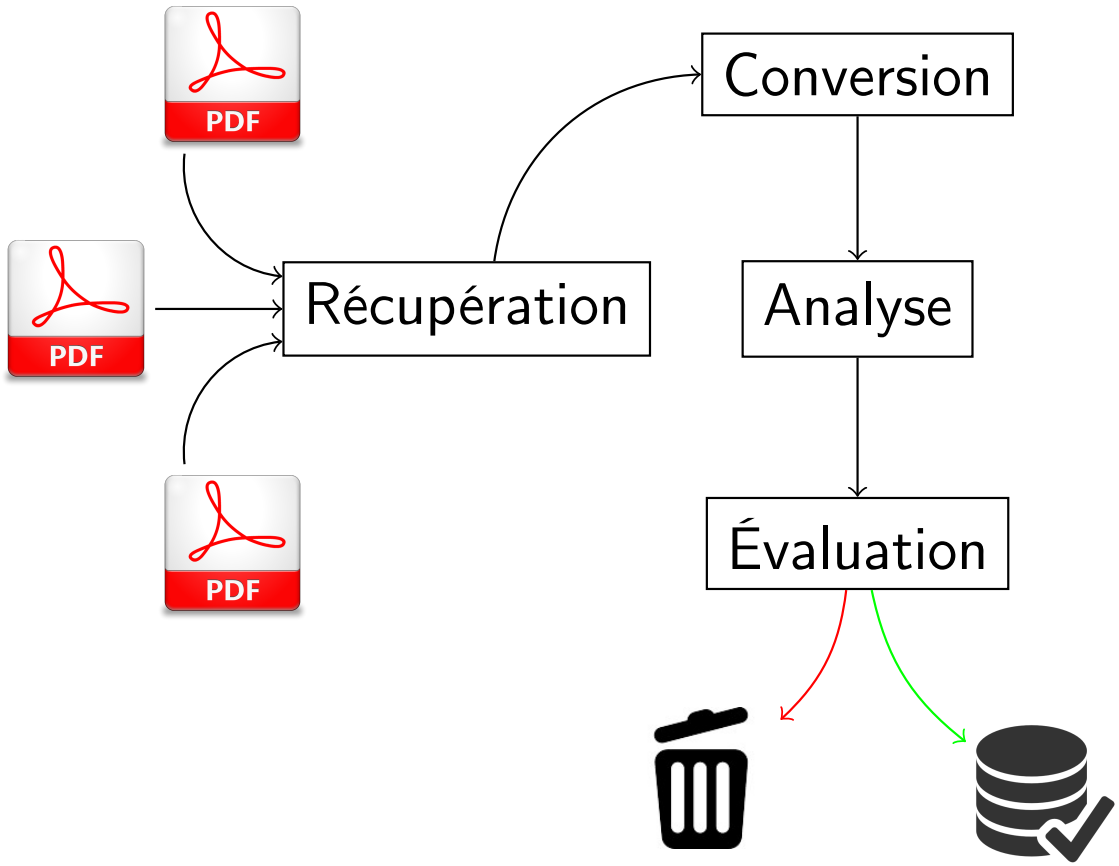
[1]ReportLinker, 21 Quai Antoine Riboud, 69002, Lyon, France
prenom.nom@reportlinker.com,
https://www.reportlinker.com/

28 Janvier 2020, Bruxelles

Plan

Introduction

Fonctionnement de ReportLinker



Fonctionnement de ReportLinker

Détection des entités nommées

Les entités nommées représentent des éléments de notre monde considérés comme uniques :

- Des personnes : *Barack Obama*, *Zeus*, *Rantanplan*
- Des lieux : *Lyon*, *Vénus*, *Asie*
- Des organisations : *Apple*, *Greenpeace*, *Fond Monétaire International*
- Des entités physiques : *Hubble*, *2CV*
- Des entités théoriques : *Réunion du G8*, *Katrina*

Traitement des informations

Plusieurs systèmes pour détecter des entités nommées :

- Traitement Automatique du Langage Naturel (NLP)
- Compréhension du Langage Naturel (NLU)

Cas des news

- Agrégateur d'actualités généralistes
- 18000 journaux ou sites web
- 4 à 6 millions de dépêches d'actualités en langue anglaise chaque mois

Problématique

- Un certain volume à traiter
- Uniquement de la donnée textuelle
- Détection de nouveaux concepts sans listes externes

Problématique

- Un certain volume à traiter
- Uniquement de la donnée textuelle
- Détection de nouveaux concepts sans listes externes

Comment générer une liste de compagnies à partir des titres de news ?

Problématique

- Un certain volume à traiter
- Uniquement de la donnée textuelle
- Détection de nouveaux concepts sans listes externes

Comment générer une liste de compagnies à partir des titres de news ?

$\Rightarrow$ utiliser des verbes d'acquisition

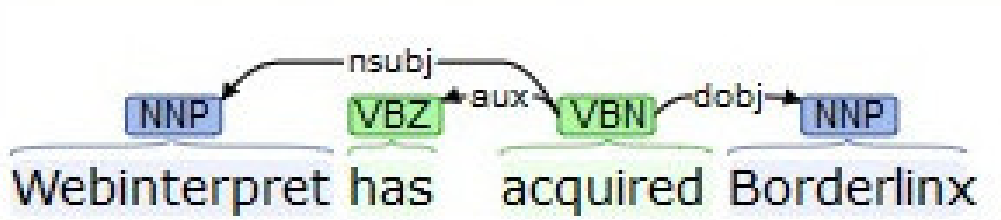
Contexte Applicatif

Modèle Stanford CoreNLP [Manning et al., 2014]

- Issue d'un travail présenté en 2014
- Version 3.9.2
- Permet un étiquetage
- Relativement lent dans sa version complète

Étiquetage standard

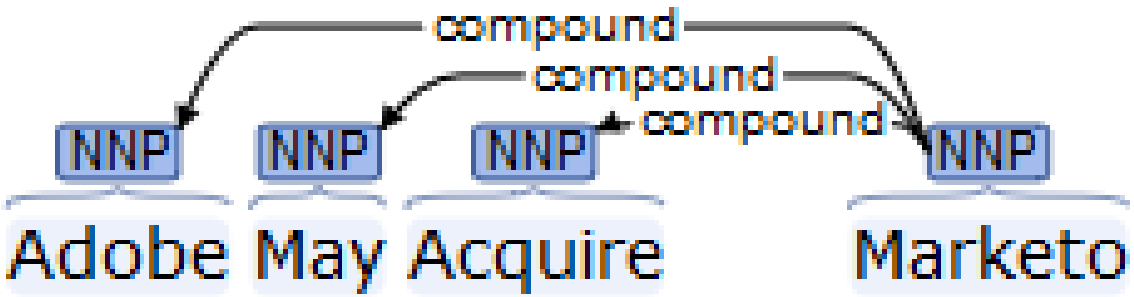
- Fonctionne sur des techniques d'apprentissages
- Permets d'expliciter les relations entre les différents éléments d'une phrase (sujet / verbe / objet)



- Permet de représenter les transactions sous la forme d'un triplet $T\langle S, VT, C\rangle$

Problèmes liés

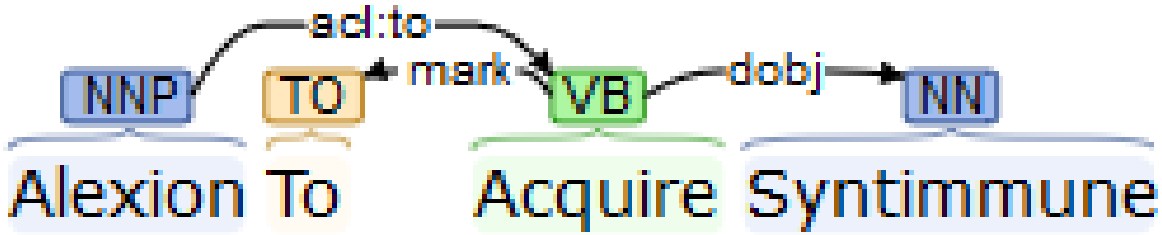
Écriture en majuscule :



- Dépendant de la mise en page du site / journal
- Trompe Standford

Problèmes liés

Style journalistique :



- Délivre de l'information essentielle en peu de mots
- Indépendamment de la langue anglaise
- N'est pas reconnu par Standford

Améliorations

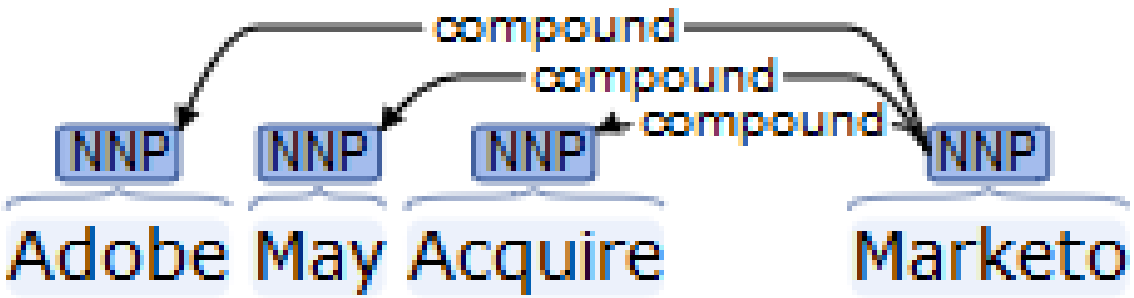
Pour améliorer le modèle de Stanford, nous avons mis en place deux pré-traitements supplémentaires portant sur :

- les majuscules
- les syntagmes verbaux VT

Majuscules

- Conservation de la majuscule sur la premiètre lettre de chaque mot
- Permet d'éviter les problèmes dût à la rédaction de titres majuscules

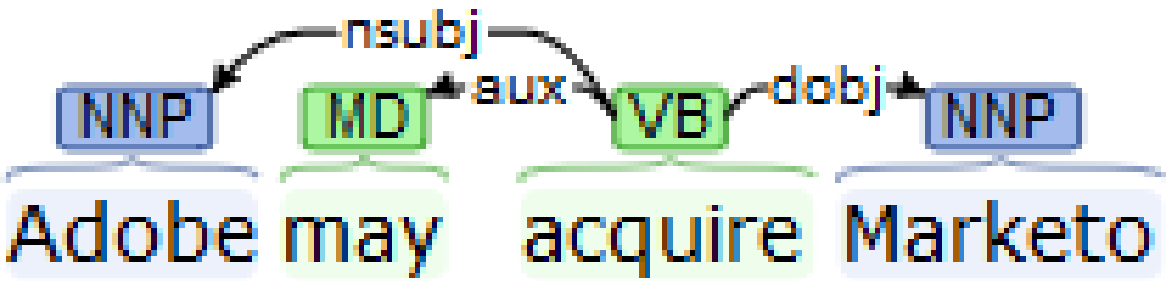
Ex :



Majuscules

- Conservation de la majuscule sur la premiètre lettre de chaque mot
- Permet d'éviter les problèmes dût à la rédaction de titres majuscules

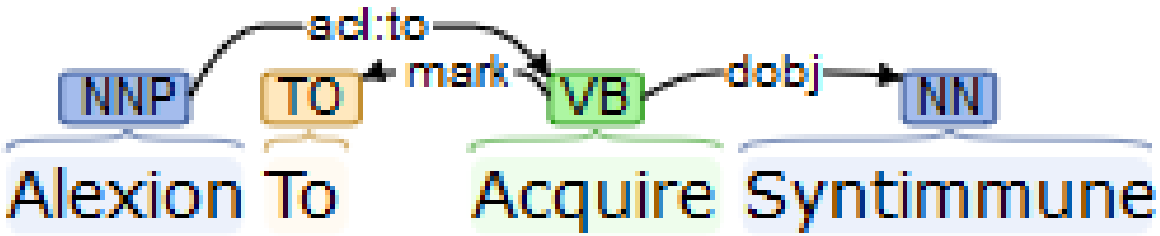
Ex :



Les syntagmes verbaux *VT*

- Non-Reconnus par Stanford
- Remplacés par des syntagmes de même sémantique
- Reconnus et étiquetés de façon exacte par le modèle amélioré

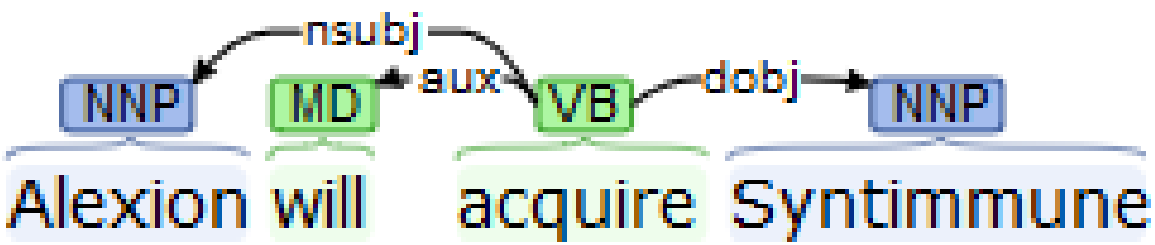
Ex :



Les syntagmes verbaux *VT*

- Non-Reconnus par Stanford
- Remplacés par des syntagmes de même sémantique
- Reconnus et étiquetés de façon exacte par le modèle amélioré

Ex :



Expérimentation

Données

Porte sur la détection de titres de news économiques en langue anglaise :

- 1500 titres avec des verbes de transactions annotés
- Dont 1253 (84%) qui comprennent deux noms de sociétés concernées par le verbe de transaction
- Règles d'annotations définies par expertise

Modèles

Nous évaluons deux modèles :

1. Le premier modèle est le modèle standard de Stanford [Manning et al., 2014], il nous sert de référence.
2. Le second modèle correspond au modèle standard auquel nous avons ajouté les deux fonctionnalités de pré-traitements.

Évaluation

Vrai positif

Un titre appartient à l'ensemble des *vrais positifs VP* s'il comporte deux sociétés qui sont détectées par le modèle.

Faux positif

Un titre appartient à l'ensemble des *faux positifs FP* s'il ne comporte pas deux sociétés mais que le modèle en détecte deux.

Faux négatif

Un titre appartient à l'ensemble des *faux négatif FN* s'il comporte deux sociétés qui ne sont pas détectées par le modèle.

Vrai négatif

Un titre appartient à l'ensemble des *vrai négatif VN* s'il ne comporte pas deux sociétés et que le modèle ne remonte pas de sociétés.

Évaluation

Précision

La *précision* offre une évaluation du bruit en calculant, parmi les titres trouvés avec deux sociétés, le pourcentage de titres qui comporte effectivement deux sociétés :

$$Précision = \frac{|VP|}{|VP| \cup |FP|} \quad (1)$$

Rappel

Le *rappel* une évaluation du silence en calculant le pourcentage de titres comprenant des sociétés trouvées parmi ceux qui sont censés l'être :

$$Rappel = \frac{|VP|}{|VP| \cup |FN|} \quad (2)$$

Évaluation

La moyenne harmonique de ces valeurs, appelée *F-mesure*, permet d'évaluer les modèles.

F-mesure

La *F-mesure* offre une évaluation des modèles :

$$F - mesure = 2 \cdot \frac{précision \cdot rappel}{précision + rappel} \tag{3}$$

Résultats

	Vrai Positif	Faux Positif	Faux Négatif	Vrai Négatif
modèle 1	719	78	534	171
modèle 2	1060	78	193	171

Table: Résultats des modèles d'extraction de noms à partir de 1500 titres de presse.

Résultats

	Distribution des trois éléments (S,VT et C)	Répartition	Pourcentage
$T1$	S ?("has|have") "acquired | bought | purchased" C	317	25,3%
$T2$	S "acquires|buys|purchases" C	169	13,5%
$T3$	C "acquired|bought|purchaded" "by" S	129	10,3%
$T4$	S "signed agreement to" "acquire|buy|purchase" C	95	7,6%
$T5$	S "will" "acquire|buy|purchase" C	9	0,7%
$T6$	S "to" "acquire|buy|purchase" C	169	13,5%
$T7$	S "acquire|buy|purchase" C	142	11,3%
Majuscule	-	30	2,4%
Autres	-	193	15,4%

Table: Présentation des règles permettant de bien classifier les titres.

Scores

	Rappel	Précision	F-mesure
modèle 1	0.57	0.90	0.70
modèle 2	0.84	0.93	0.88

Table: Scores des modèles d'extraction de noms d'entreprises à partir de titres de presse.

Conclusion et ouvertures

Résumé

- Détection automatique d'entités nommées se rapportant à des entreprises dans des titres de news
- Amélioration de la détection standard via :
  - la prise en compte de l'écriture des titres en majuscules
  - la définition de syntagmes verbaux adaptés aux entités recherchées
- L'expérimentation montre un gain dans la détection des sociétés
- Permet d'obtenir des relations typées *transaction* de qualité entre deux sociétés

Discussion

- Pas de comparaisons avec d'autres modèles
- Pas de gains sur le bruit produit
- Étape contraignante

Ouvertures

- Comparaison avec des méthodes d'apprentissage, ...
- Prise en compte de termes déclencheurs comme les formes juridiques ("inc", "ltd", etc.)
- Utilisation du corps de l'article

Références

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014).
The stanford corenlp natural language processing toolkit.
In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Fin

Merci

Des questions ?