Concilier les triplets extraits à partir de textes et ceux extraits de bases de connaissances

David ALFONSO HERMELO

Université Paris III Sorbonne Nouvelle

david.alfonso.hermelo@gmail.com

12 septembre 2017

Présentation

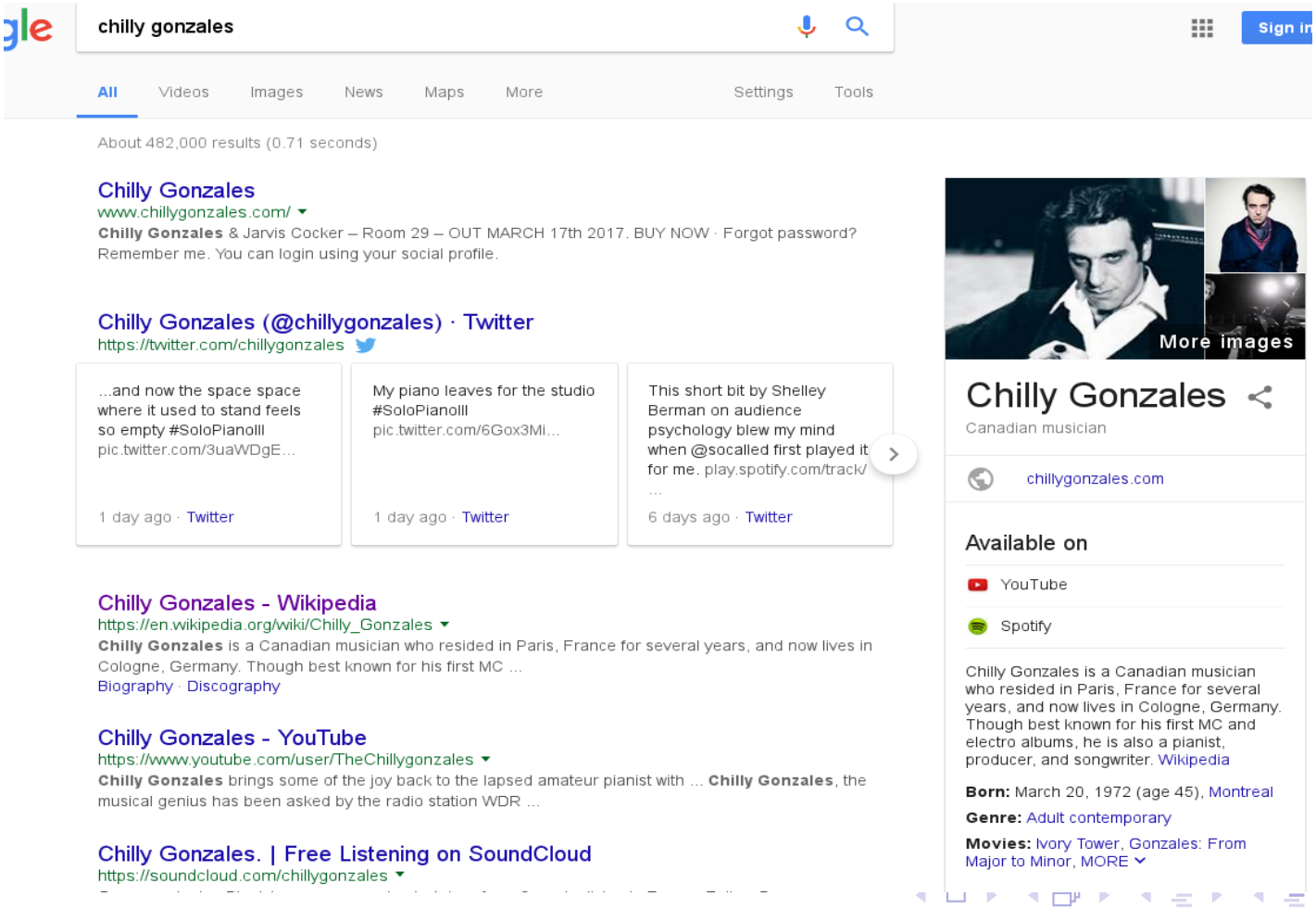
1. Introduction
   - Motivation
2. Développement
   - État de l'art et travaux connexes
   - Bases de connaissances
   - Extracteurs d'information libre
   - Analyses
   - Analyses
   - Analyses
   - Résultats
3. Conclusions
   - Contributions
   - Perspectives

Introduction

Introduction

KB (Base de Connaissances):

- Collection de faits attestés (objectifs).



Introduction

OIE (Extracteur d'information libre):

- Collection de triplets non attestés (subjectifs).

INPUT:

> Chilly Gonzales is a Canadian musician who resided in Paris for several years, and now lives in Cologne (which is in Germany).

OUTPUT:

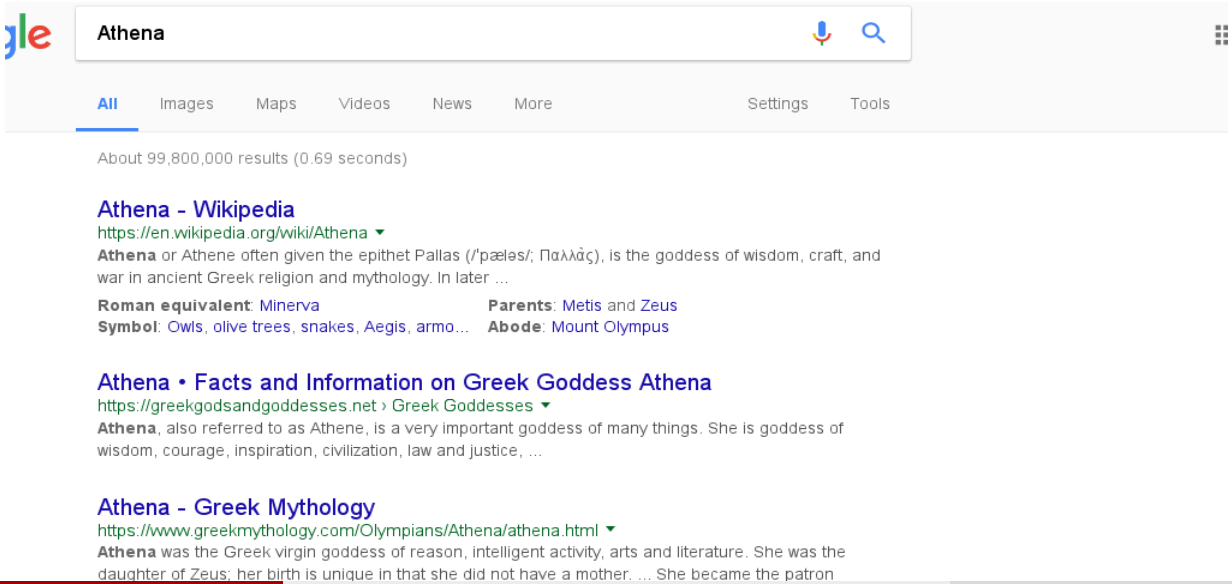
> Chilly Gonzales : is a : Canadian musician
> Gonzales : is : Canadian
> Cologne : is : in Germany
> now lives in Cologne : resided : in Paris for several years
> Paris : now lives : in Cologne
> which : is : in Germany

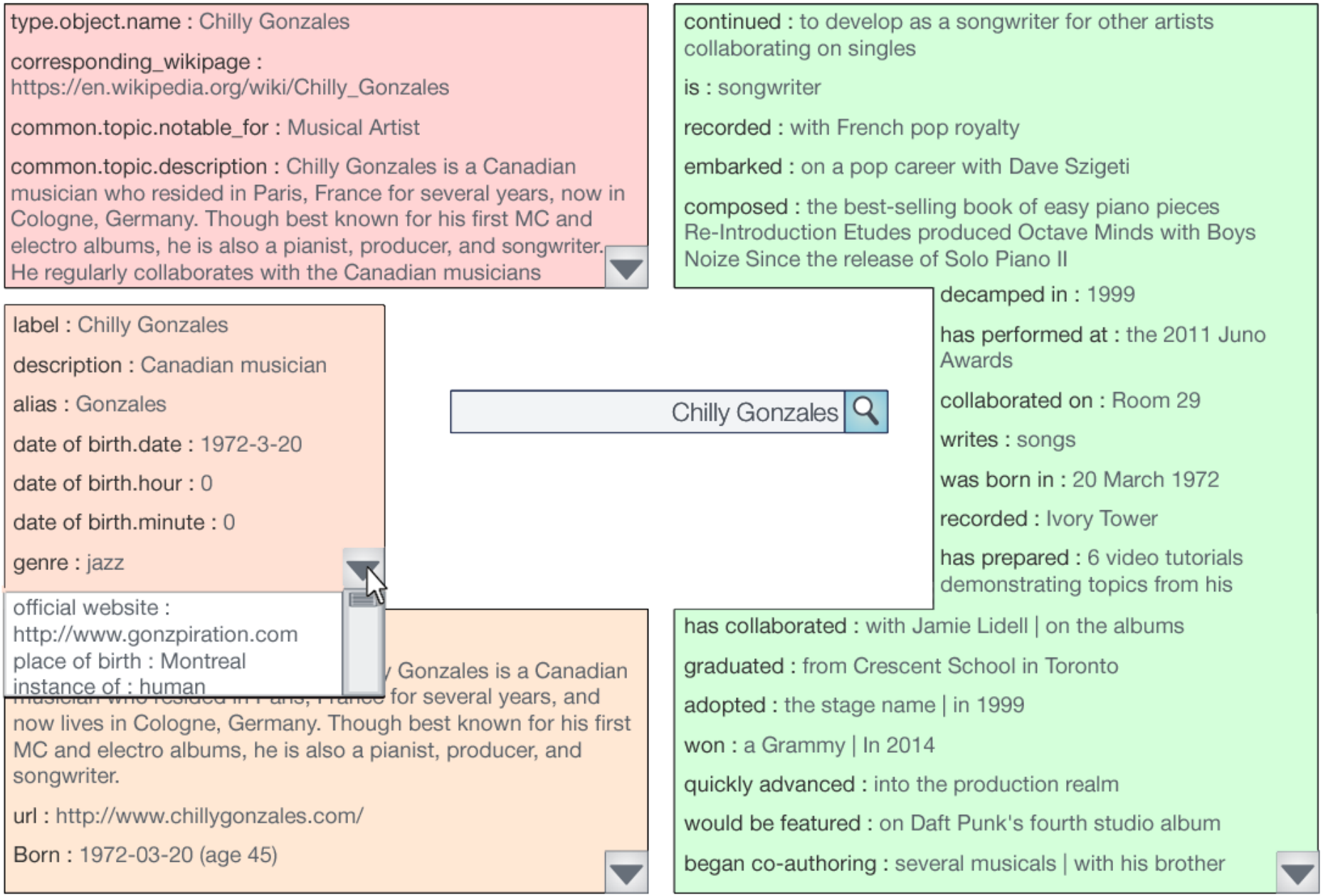
Motivation

- Analyse de la complémentarité des KB et des OIE.

Pourquoi?

- Manque de mesures.
- Les informations subjectives peuvent être utiles:
  - Les vins italiens sont les meilleurs au monde.
  - Les vins français sont les meilleurs au monde.
- Les KB ont des lacunes informatives:




type.object.name : Chilly Gonzales
corresponding_wikipage : https://en.wikipedia.org/wiki/Chilly_Gonzales
common.topic.notable_for : Musical Artist
common.topic.description : Chilly Gonzales is a Canadian musician who resided in Paris, France for several years, now in Cologne, Germany. Though best known for his first MC and electro albums, he is also a pianist, producer, and songwriter. He regularly collaborates with the Canadian musicians
continued : to develop as a songwriter for other artists collaborating on singles
is : songwriter
recorded : with French pop royalty
embarked : on a pop career with Dave Szigeti
composed : the best-selling book of easy piano pieces Re-Introduction Etudes produced Octave Minds with Boys Noize Since the release of Solo Piano II
label : Chilly Gonzales
description : Canadian musician
alias : Gonzales
date of birth.date : 1972-3-20
date of birth.hour : 0
date of birth.minute : 0
genre : jazz
official website : http://www.gonzpiration.com
place of birth : Montreal
instance of : human
Chilly Gonzales
decamped in : 1999
has performed at : the 2011 Juno Awards
collaborated on : Room 29
writes : songs
was born in : 20 March 1972
recorded : Ivory Tower
has prepared : 6 video tutorials demonstrating topics from his
Gonzales is a Canadian
for several years, and
now lives in Cologne, Germany. Though best known for his first MC and electro albums, he is also a pianist, producer, and songwriter.
url : http://www.chillygonzales.com/
Born : 1972-03-20 (age 45)
has collaborated : with Jamie Lidell | on the albums
graduated : from Crescent School in Toronto
adopted : the stage name | in 1999
won : a Grammy | In 2014
quickly advanced : into the production realm
would be featured : on Daft Punk's fourth studio album
began co-authoring : several musicals | with his brother

Développement

ÉTAPES	MÉTHODES
• Analyse de l'état de l'art et travaux connexes.	• Observation empirique.

Projets KB analysés

Yago
Yago2
DBpedia
Nell
Wikidata
Freebase
Knowledge Graph
Knowledge Vault

ÉTAPES	MÉTHODES
• Analyse de l'état de l'art et travaux connexes.	• Observation empirique.

Projets OIE analysés

ReVerb
OLLIE
ClausIE
PropS
CSD-IE
OpenIE-4
Standford Open IE

ÉTAPES	MÉTHODES
• Sélection des KB les plus complètes.	• Observation empirique. • Analyse statistique.

	Nb entités (millions)	Nb faits (millions)
YAGO	10	120
NELL	2.98	90
DBpedia	4	580
Deep Dive	0.055	0.980
Wikidata	**25**	**144**
Freebase	**58**	**1900**
Google Knowledge Graph	**1000**	**70000**

Table: Taille approximative de plusieurs KB

ÉTAPES	MÉTHODES
• Choix des entités pour analyses.	• Observation empirique.

Procédure

- 100 entités articles Wikipédia au hasard + 1 entité fétiche.
- Suppression des entités $\neg \exists$ dans les 3 KB les plus complètes.
- 20 entités = requêtes pour analyses.

Freebase	Know. Graph	Wikidata
Chilly Gonzales	Chilly Gonzales	Chilly Gonzales
Hoja en Blanco	Hoja en Blanco	Hoja en Blanco
Copelatus unguicularis	Copelatus unguicularis	Copelatus unguicularis
Ajaigarh State	Ajaigarh State	Ajaigarh State
Clare Mill, California	Clare Mill, California	Clare Mill, California
Hellboy: Blood and Iron	Hellboy: Blood and Iron	Hellboy: Blood and Iron
Future Legends	Future Legends	Future Legends
McAlmont & Butler	McAlmont & Butler	McAlmont & Butler
1987 Rose Bowl	1987 Rose Bowl	1987 Rose Bowl
Taranaki rugby league team	Taranaki rugby league team	Taranaki rugby league team

ÉTAPES	MÉTHODES
• Sélection des OIE plus compétents.	• Analyse statistique.

Schéma du banc d'essai extrait de Stanovsky et Dagan (2016).

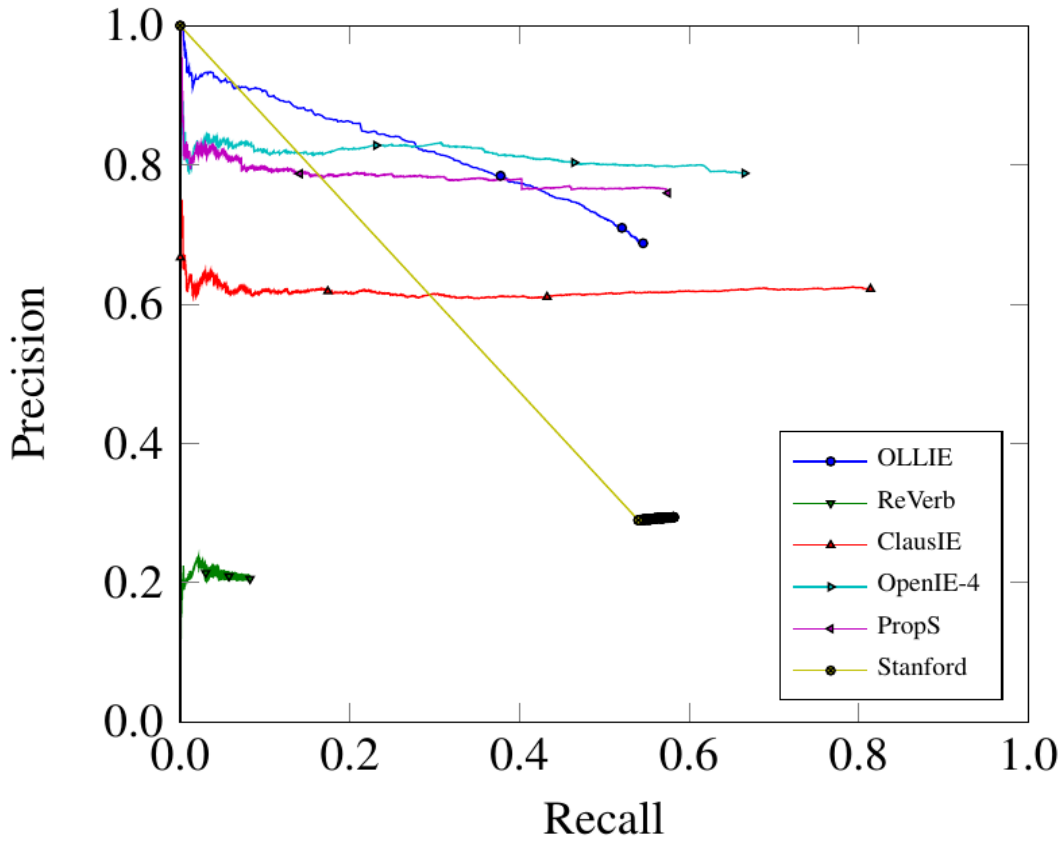


Schéma Wikipédia.

ÉTAPES	MÉTHODES
• Analyse des KB sélectionnées.	• Analyse statistique.

Nous avons compté les faits communs à plusieurs KB.

Choix correspondance S-O ("match"):

Freebase	Chilly Gonzales	music.artist.origin	Montreal
Wikidata	Chilly Gonzales	place of birth	Montreal

Freebase	Chilly Gonzales	music.artist.origin	Montreal
Wikidata	Chilly Gonzales	place of birth	Canada

Analyse des KB sélectionnées

	Freebase	Know. Graph	Wikidata	TOTAL FAITS
Freebase	-	63	60	770
Know. Graph	-	-	39	250
Wikidata	-	-	-	394
TOTAL FAITS	770	250	394	

Table: Tableau de contingence du nombre de faits en commun pour les différentes KB (des 20 E d'échantillon).

ÉTAPES	MÉTHODES
• Analyse des triplets de sortie des OIE sélectionnés.	• Analyse statistique.

Nous avons compté les triplets communs à plusieurs OIE.

Choix correspondance S-P-O ("match"):

Phrase 1 - ClausIE	Chilly Gonzales	was born in	Montreal
Phrase 117 - OLLIE	Chilly Gonzales	was born in	Montreal

Phrase 1 - ClausIE	Chilly Gonzales	was born in	Montreal
Phrase 225 - OLLIE	Chilly Gonzales	was born at	Montreal

Phrase 1 - ClausIE	Chilly Gonzales	was born in	Montreal
Phrase 854 - OLLIE	Chilly Gonzales	played at	Montreal

Phrase 45 - ClausIE	His brother	played at	the same concert
Phrase 854 - OLLIE	Chilly Gonzales	played at	Montreal

Analyse des OIE sélectionnées

	ClausIE	OLLIE	OpenIE-4	TOTAL
ClausIE	-	344	583	10501
OLLIE	-	-	0	8331
OpenIE-4	-	-	-	6729
TOTAL	10501	8331	6729	

Table: Tableau de contingence du nombre de triplets en commun en sortie des différentes OIE (des 20 E d'échantillon, avec le contenu des 10 premiers résultats de Google Search + Wikipédia pour entrée d'OIE).

Étapes et méthodes de recherche

ÉTAPES	MÉTHODES
• Expérimentation comparative triplets de sortie OIE - faits de KB.	• Recherche expérimentale. • Analyse statistique.

	Freebase	Knowledge Graph	Wikidata	TOTAL
ClausIE	4	1	6	10501
OLLIE	0	0	0	8331
OpenIE-4	2	3	4	6729
TOTAL	765	153	404	

Table: ableau de contingence du nombre de triplets en commun en sortie des différentes OIE et KB (des 20 E d'échantillon).

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word emb.
- Sujet alias
- Sujet pronom

• Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign • he: had re-signed : with a major label • it: drew : immediate comparisons to the work of Erik Satie • Gonzales: wrote : music • everyone: to sing : for me • livre de partitions: à : jouer • he: would win : a Grammy Award • Gonzales: was born : on 20 March 1972 • Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours	• the artist: took up : residence in Europe mainly in Germany just after signing to Kitty-Yo • he: had re-signed : with a major label • it: drew : immediate comparisons to the work of Erik Satie • he: is : songwriter • Gonzo: shares : his point of view • écrit des: chansons : avec Jarvis Cocker • Gonzales: was born : on 20 March 1972 • Gonzales also known as Chilly Gonzales: is : an MC-meets-keyboarding-producer-meets-singer-extraordinaire

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word emb.
- Sujet alias
- Sujet pronom

• Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign • he: had re-signed : with a major label • it: drew : immediate comparisons to the work of Erik Satie • Gonzales: wrote : music • everyone: to sing : for me • livre de partitions: à : jouer • he: would win : a Grammy Award • Gonzales: was born : on 20 March 1972 • Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours	• the artist: took up : residence in Europe mainly in Germany just after signing to Kitty-Yo • he: had re-signed : with a major label • it: drew : immediate comparisons to the work of Erik Satie • he: is : songwriter • Gonzo: shares : his point of view • écrit des: chansons : avec Jarvis Cocker • Gonzales: was born : on 20 March 1972 • Gonzales also known as Chilly Gonzales: is : an MC-meets-keyboarding-producer-meets-singer-extraordinaire

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

0,92 - Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign
0,41 - he: had re-signed : with a major label
0,27 - it: drew : immediate comparisons to the work of Erik Satie
0,92 - Gonzales: wrote : music
0,78 - everyone: to sing : for me
0,92 - livre de partitions: à : jouer
0,50 - he: would win : a Grammy Award
0,93 - Gonzales: was born : on 20 March 1972
0,97 - Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours
0,92 - Gonzo: shares : his point of view on scales

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

0,92 - Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign
0,41 - he: had re-signed : with a major label
0,27 - it: drew : immediate comparisons to the work of Erik Satie
0,92 - Gonzales: wrote : music
0,78 - everyone: to sing : for me
0,92 - livre de partitions: à : jouer
0,50 - he: would win : a Grammy Award
0,93 - Gonzales: was born : on 20 March 1972
0,97 - Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours
0,92 - Gonzo: shares : his point of view on scales

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

0,92 - Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign
0,41 - he: had re-signed : with a major label
0,27 - it: drew : immediate comparisons to the work of Erik Satie
0,92 - Gonzales: wrote : music
0,78 - everyone: to sing : for me
0,92 - livre de partitions: à : jouer
0,50 - he: would win : a Grammy Award
0,93 - Gonzales: was born : on 20 March 1972
0,97 - Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours
0,92 - Gonzo: shares : his point of view on scales

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

0,92 - Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign
0,41 - he: had re-signed : with a major label
0,27 - it: drew : immediate comparisons to the work of Erik Satie
0,92 - Gonzales: wrote : music
0,78 - everyone: to sing : for me
0,92 - livre de partitions: à : jouer
0,50 - he: would win : a Grammy Award
0,93 - Gonzales: was born : on 20 March 1972
0,97 - Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours
0,92 - Gonzo: shares : his point of view on scales

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

0,92 - Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign
0,41 - he: had re-signed : with a major label
0,27 - it: drew : immediate comparisons to the work of Erik Satie
0,92 - Gonzales: wrote : music
0,78 - everyone: to sing : for me
0,92 - livre de partitions: à : jouer
0,50 - he: would win : a Grammy Award
0,93 - Gonzales: was born : on 20 March 1972
0,97 - Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours
0,92 - Gonzo: shares : his point of view on scales

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

0,92 - Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign
0,41 - he: had re-signed : with a major label
0,27 - it: drew : immediate comparisons to the work of Erik Satie
0,92 - Gonzales: wrote : music
0,78 - everyone: to sing : for me
0,92 - livre de partitions: à : jouer
0,50 - he: would win : a Grammy Award
0,93 - Gonzales: was born : on 20 March 1972
0,97 - Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours
0,92 - Gonzo: shares : his point of view on scales

ÉTAPES	MÉTHODES
• Conceptualisation d'heuristiques.	• Observation empirique.

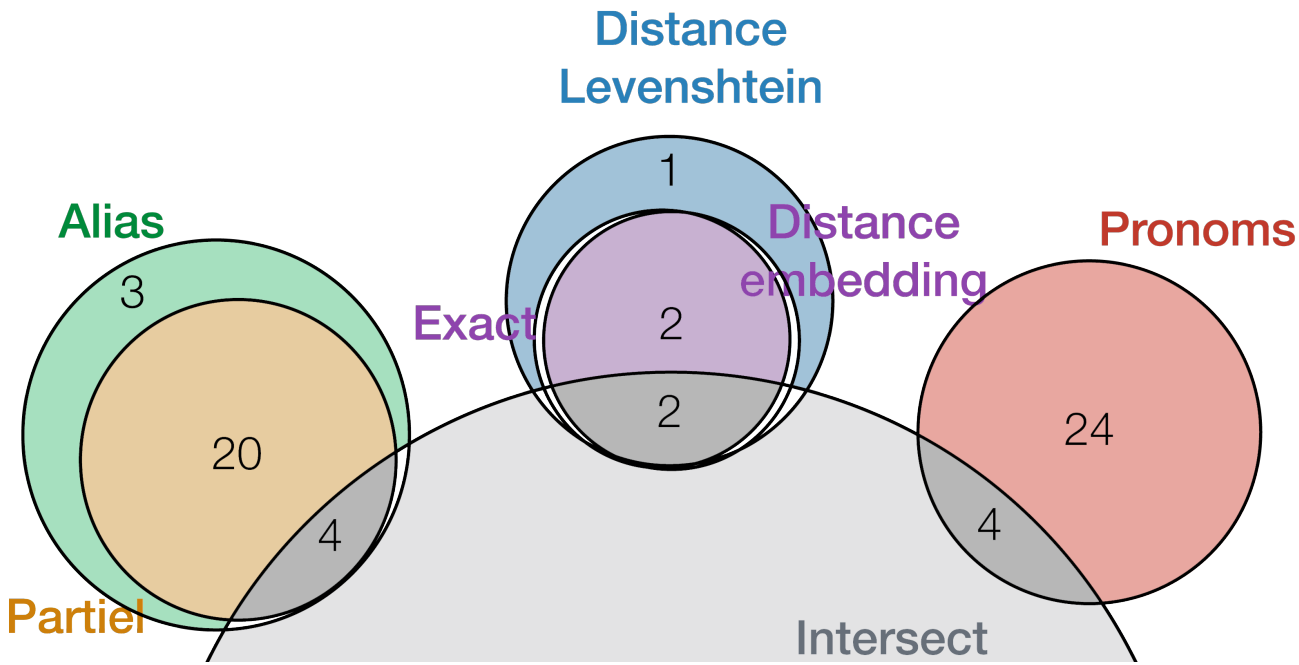
Heuristiques

- Union et intersection
- Score de confiance
- Sujet exact
- Sujet entité-partielle
- Sujet dist. Levenshtein
- Sujet dist. word embedding
- Sujet alias
- Sujet pronom

0,92 - Chilly Gonzales: composed : a global hit for the inaugural Apple iPad campaign
0,41 - he: had re-signed : with a major label
0,27 - it: drew : immediate comparisons to the work of Erik Satie
0,92 - Gonzales: wrote : music
0,78 - everyone: to sing : for me
0,92 - livre de partitions: à : jouer
0,50 - he: would win : a Grammy Award
0,93 - Gonzales: was born : on 20 March 1972
0,97 - Chilly Gonzales: holds : the Guinness world record for the longest solo concert at over 27 hours
0,92 - Gonzo: shares : his point of view on scales

ÉTAPES	MÉTHODES
• Expérimentations quantitatives des heuristiques.	• Recherche expérimentale. • Analyse statistique.

Diagramme Venn des triplets de sortie OpenIE-4 capturés par les heuristiques pour l'entité Chilly Gonzales.



Catégorie	Total
Exact	4
Distance Levenshtein	5
Distance embedding (word2vec)	4
Alias	27
Partiel	24
Pronoms	28
Intersection OpenIE-4 & ClausIE	9

Schéma du mémoire 1. Schéma du mémoire 2.

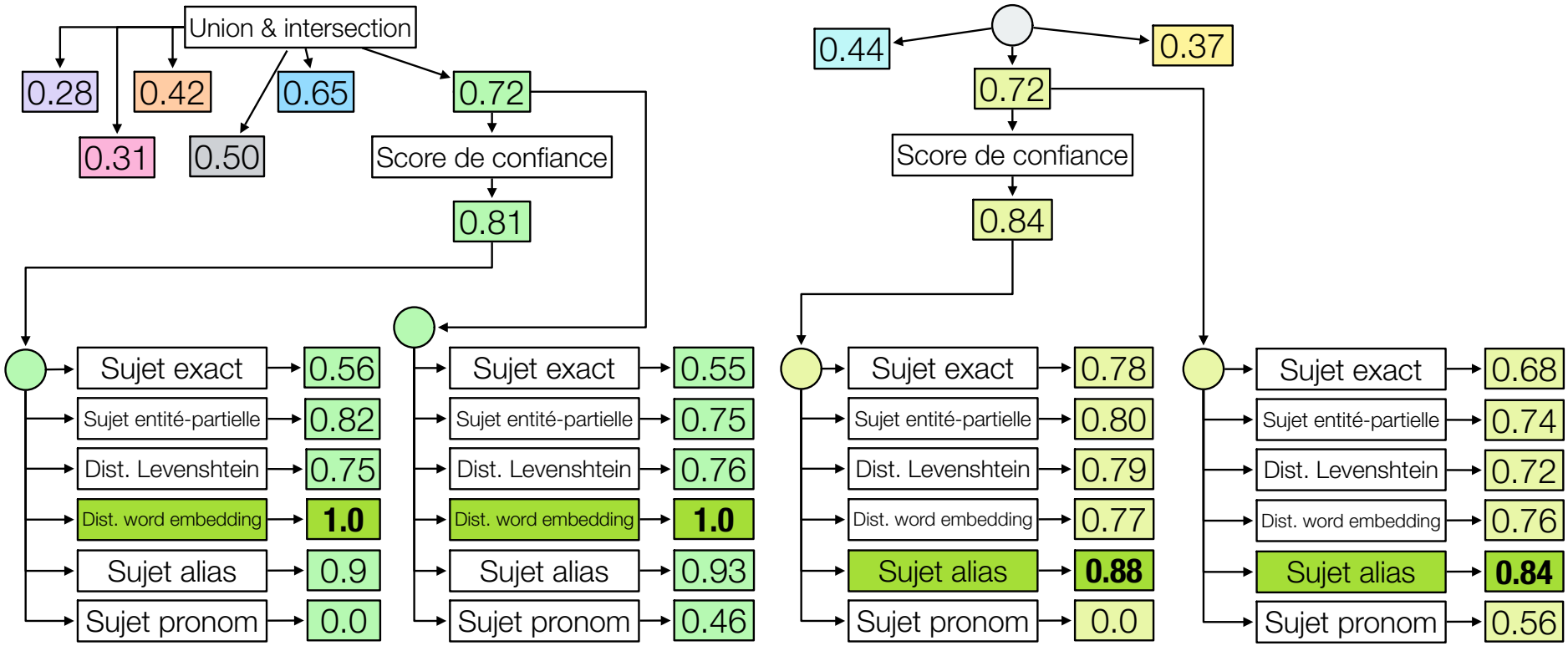
ÉTAPES

- Expérimentations qualitatives des heuristiques.

MÉTHODES

- Recherche expérimentale.
- Analyse statistique.

Scores de précision des heuristiques (banc d'essai Stanovsky et Dagan (2016))



Résultats

ÉTAPES	MÉTHODES
• Analyse des résultats des expériences.	• Analyse statistique.

Triplets d'entrée	Heuristique	Score précision	Score Rappel
OpenIE-4	S proche distance Embedding	0.76	0.0006
OpenIE-4	S alias de E	0.84	0.004
OpenIE-4	S proche distance Embedding $\cap$ Score de confiance $>$ 0.9	0.77	0.001
OpenIE-4	S alias de E $\cap$ Score de confiance $>$ 0.9	0.88	0.003
ClausIE $\cap$ OpenIE-4	S proche distance Embedding	1.0	0.0001
ClausIE $\cap$ OpenIE-4	S alias de E	0.93	0.001
ClausIE $\cap$ OpenIE-4	S proche distance Embedding $\cap$ Score de confiance $>$ 0.9	1.0	0.0001
ClausIE $\cap$ OpenIE-4	S alias de E $\cap$ Score de confiance $>$ 0.9	0.9	0.0004

Conclusions

Contributions

- Analyse comparative des complémentarités (strictes) entre:
  - des faits de KB choisies.
  - des triplets de sortie d'OIE choisis.
  - des faits de KB choisies et des triplets de sortie d'OIE choisis
- Évaluation d'heuristiques appliquées sur des triplets de sortie OIE.

Perspectives

- Approfondir dans l'analyse statistique entre faits de KB et triplets de sortie OIE.
- Concevoir un banc d'essai plus complet.
- Poursuivre la recherche de meilleures heuristiques ou métaheuristiques.
- Tester l'efficacité des réseaux de neuronnes ou des modèles statistiques vs. l'efficacité des heuristiques.
- Concevoir, développer et tester une KB à faits subjectifs.

Références capitales

Xin Dong et al. (2014)
Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion
Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 601 – 610.

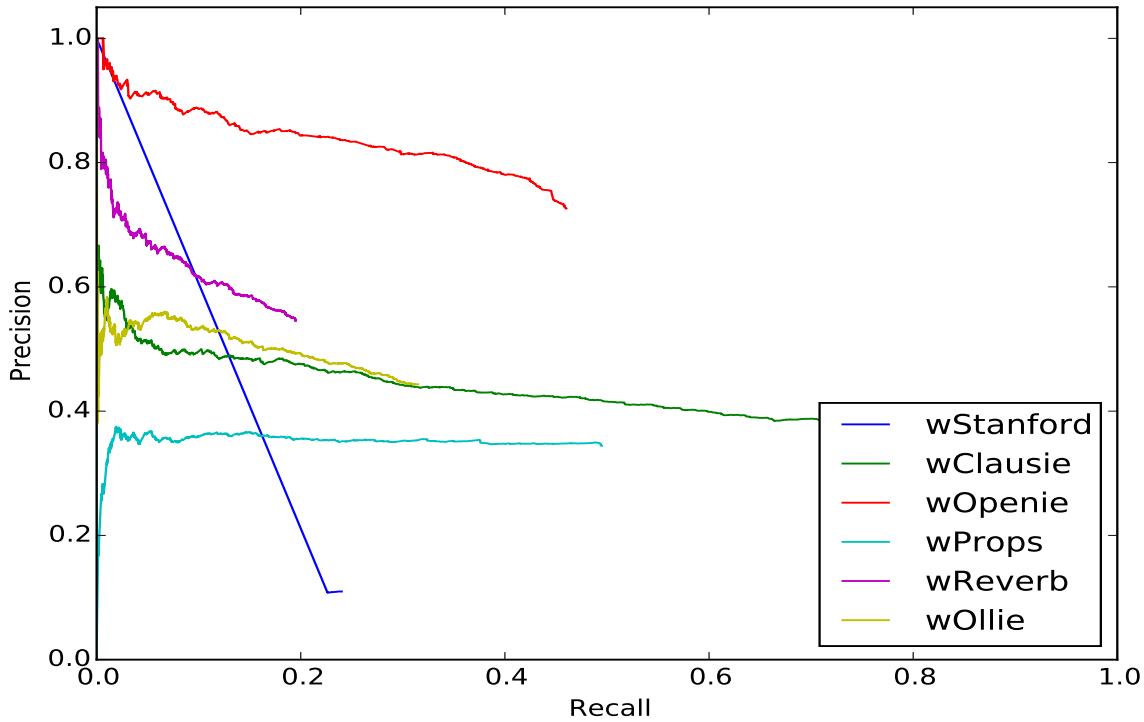
Gabriel Stanovsky et Ido Dagan (2016)
Creating a large benchmark for open information extraction
Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Mausam (2016)
Open Information Extraction Systems and Downstream Applications
Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 4074 – 4077.

Questions

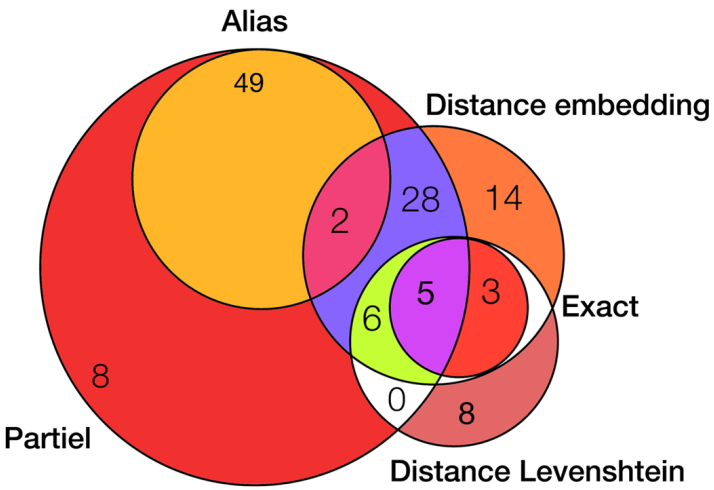
Schema du mémoire.

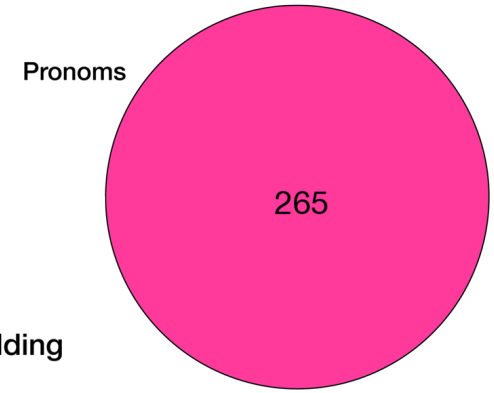
Schéma du banc d'essai de Stanovsky et Dagan (2016) en n'utilisant que le corpus extrait de Wikipédia.



Schema du mémoire.

Diagramme de Venn - Corpus tiré de Google Search - outil OpenIE-4

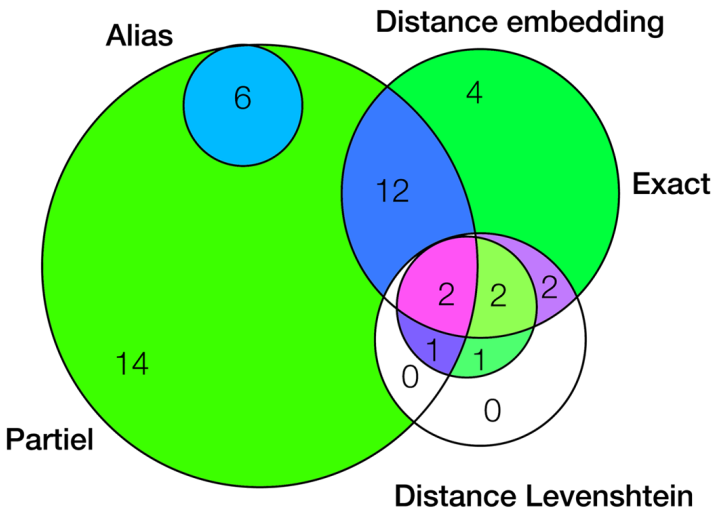


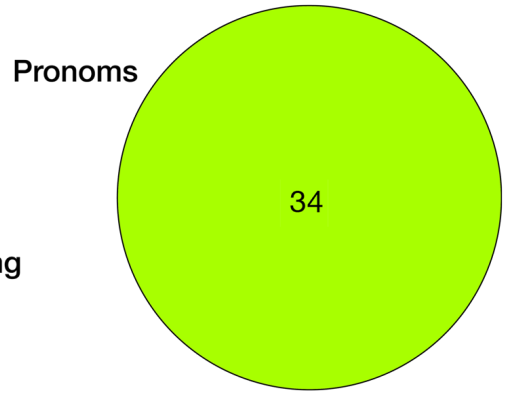


Catégorie	Total
Exact	8
Distance Levenshtein	22
Distance embedding (word2vec)	58
Alias	51
Partiel	92
Pronoms	265

Schema du mémoire.

Diagramme de Venn - Corpus tiré de Wikipédia - outil OpenIE-4





Catégorie	Total
Exact	6
Distance Levenshtein	8
Distance embedding (word2vec)	22
Alias	6
Partiel	23
Pronoms	34