Introduction à l'extraction d'information

felipe@iro.umontreal.ca

RALI
Dept. Informatique et Recherche Opérationnelle
Université de **Montréal**

mars 2015, V0.1

Plan

Extracteurs

Structuration

Applications

Analyse

Datasets

rali

Extraction d'information

But : extraire de l'information structurée depuis des documents (principalement) non structurés

- les premiers systèmes étaient dédiés à certains domaines et certaines relations
  find(Companies,acquired,Companies)
  - règles spécifiques (reg exp) et **supervision distante**
- Open Information Extraction
  - pionnier : TEXTRUNNER ***[Banko et al., 2007]***
  - acquisition de toute relation
  - sans supervision (ou presque)
  - en utilisant de grandes quantités de textes

rali

Pour quoi faire ?

- Web Sémantique
- Common sense knowledge (le rêve de l'IA classique est de retour)
- Extraction automatique d'ontologies
- Meilleure recherche d'information (ex : GOOGLE's Knowledge Vault)
- Question answering (regain d'intérêt depuis IBM Watson)

- En fait, utile à toute application du TALN d'intérêt !

rali

Plan

TEXTRUNNER

http://www.cs.washington.edu

- Le premier extracteur ouvert
  OIE = Open Information Extraction
- Les autres extracteurs sont fortement inspirés de celui-ci
- Les problèmes d'évaluation sont bien abordés

- Lire *[Banko et al., 2007]*

rali

TEXTRUNNER

1. Un classificateur de triplets (pertinent/non pertinent)
   - auto-entraînement sur quelques milliers de phrases du PTB
   - traits du type : POS dans les relations, dans les arguments (avant/après, etc.)
   - modèle naive-bayes (dans l'article), CRF par la suite
2. Un extracteur basé sur les POS
   - noun-phase chunker (modèle *maxent*)
   - relations identifiées en normalisant le texte entre deux groupes nominaux (*definitely developed* $\Rightarrow$ *developed*)
   - consultation du classificateur (filtre)
3. Filtre de redondance
   - utilise le nombre de phrases différentes supportant un triplet (normalisé) pour décider de la pertinence du triplet
   - selon un modèle $p(\text{pertinence}|\text{nb. of support sentences})$

rali

TEXTRUNNER

- Extraction sur ClueWeb
  - 9M pages = 133M phrases = 60.5M tuples (2.2 tuples par phrase)
  - 0.65 CPU seconds par page Web
    68 CPU hours pour traiter ClueWeb sur une seule machine
- Évaluation :
  1. comparé à KNOWITALL sur 10 relations supportées au moins 1000 fois dans un corpus de 9M de phrases
     $\{<\textit{nom-propre}>, \textit{acquired}, <\textit{nom-propre}>\}$,
     $\{<\textit{nom-propre}>, \textit{graduated-from}, <\textit{nom-propre}>\}$, ...
     $\Rightarrow$ environ la même performance (pas de détail par relation)
  2. Évaluation des triplets impliquant les autres relations
     - 11.3 M de triplets avec prob ≥ 0.8, supportés par au moins 10 phrases et pas une relation dans le top 0.1% (*has*, *is*)

TEXTRUNNER

- Extraction sur ClueWeb
  - 9M pages = 133M phrases = 60.5M tuples (2.2 tuples par phrase)
  - 0.65 CPU seconds par page Web
    68 CPU hours pour traiter ClueWeb sur une seule machine
- Évaluation :
  1. comparé à KNOWITALL sur 10 relations supportées au moins 1000 fois dans un corpus de 9M de phrases
     $\{<\textit{nom-propre}>, \textit{acquired}, <\textit{nom-propre}>\}$,
     $\{<\textit{nom-propre}>, \textit{graduated-from}, <\textit{nom-propre}>\}$, ...
     $\Rightarrow$ environ la même performance (pas de détail par relation)
  2. Évaluation des triplets impliquant les autres relations
     - 11.3 M de triplets avec prob ≥ 0.8, supportés par au moins 10 phrases et pas une relation dans le top 0.1% (*has*, *is*)
     - 7.8M sont bien formés (c-ex : $\{\textit{29}, \textit{dropped}, \textit{instruments}\}$)

TEXTRUNNER

- Extraction sur ClueWeb
  - 9M pages = 133M phrases = 60.5M tuples (2.2 tuples par phrase)
  - 0.65 CPU seconds par page Web
    68 CPU hours pour traiter ClueWeb sur une seule machine
- Évaluation :
  1. comparé à KNOWITALL sur 10 relations supportées au moins 1000 fois dans un corpus de 9M de phrases
     $\{<\textit{nom-propre}>, \textit{acquired}, <\textit{nom-propre}>\}$,
     $\{<\textit{nom-propre}>, \textit{graduated-from}, <\textit{nom-propre}>\}$, ...
     $\Rightarrow$ environ la même performance (pas de détail par relation)
  2. Évaluation des triplets impliquant les autres relations
     - 11.3 M de triplets avec prob ≥ 0.8, supportés par au moins 10 phrases et pas une relation dans le top 0.1% (*has*, *is*)
     - 7.8M sont bien formés (c-ex : $\{\textit{29}, \textit{dropped}, \textit{instruments}\}$)
     - évaluation manuelle d'un échantillon de 400 triplets : $\sim$80% de triplets OK (parmi ceux correctement formés)

rali

TEXTRUNNER

- Extraction sur ClueWeb
  - 9M pages = 133M phrases = 60.5M tuples (2.2 tuples par phrase)
  - 0.65 CPU seconds par page Web
    68 CPU hours pour traiter ClueWeb sur une seule machine
- Évaluation :
  1. comparé à KNOWITALL sur 10 relations supportées au moins 1000 fois dans un corpus de 9M de phrases
     $\{<\textit{nom-propre}>, \textit{acquired}, <\textit{nom-propre}>\}$,
     $\{<\textit{nom-propre}>, \textit{graduated-from}, <\textit{nom-propre}>\}$, ...
     $\Rightarrow$ environ la même performance (pas de détail par relation)
  2. Évaluation des triplets impliquant les autres relations
     - 11.3 M de triplets avec prob $\geq$ 0.8, supportés par au moins 10 phrases et pas une relation dans le top 0.1% (*has*, *is*)
     - 7.8M sont bien formés (c-ex : $\{\textit{29}, \textit{dropped}, \textit{instruments}\}$)
     - évaluation manuelle d'un échantillon de 400 triplets : $\sim$80% de triplets OK (parmi ceux correctement formés)
     - évaluation manuelle concret (ex : $\{\textit{Telsa}, \textit{invented}, \textit{coli tranform}\}$) vs. abstrait (ex : $\{\textit{executive}, \textit{hired-by}, \textit{company}\}$) — 14% - 86%

TEXTRUNNER

À propos du classificateur auto-entrainé . . .

- analyse en dépendances des phrases du corpus ”d’entraînement”
- identification de gpes nominaux non récurrents (chunks) e_1 et e_2
- extraction d’une relation candidate en parcourant le graphe de dépendances
- étiquetage comme exemple positif si :
  1. chemin entre e_1 et e_2 pas trop long (constante)
  2. ne croisant pas de frontières (comme les clauses relatives)
  3. e_1 et e_2 ne contiennent pas de pronoms
- les exemples sont ensuite encodés (features) et donnés à un apprenant (naive-bayes)

rali

WOE

- **idée :** utiliser les infobox de WIKIPEDIA comme supervision pour entraîner un extracteur

Pierre Lapointe

Pour les articles homonymes, voir Lapointe.

Pierre Lapointe (né le 23 mai 1981 à Alma, au Québec) est un auteur-compositeur-interprète canadien (québécois).

Son œuvre s'inscrit d'une part dans la tradition de la chanson francophone, qu'elle soit québécoise ou française, ce qui se vérifie dans des textes souvent très littéraires[non neutre]. Pierre Lapointe est également influencé par la musique pop qu'il entend utiliser pour renouveler la première citée[1]. Les arts graphiques contemporains, en particulier l'art numérique, sont très présents dans son univers via ses vidéoclips, et participent à la création d'un univers onirique et paradoxal, entre chansons mélancoliques et obscures, et scénographies colorées voire provocatrices.

Se définissant lui-même comme « chanteur populaire »[2,3], il s'est construit un personnage de dandy égocentrique de l'élite artistique montréalaise[4], lui permettant d'imprimer un décalage volontaire entre l'artiste sur scène et sa production largement biographique.

Ses disques enregistrent un succès critique et commercial au Canada[4].

Sommaire [masquer]

1 Biographie
- 1.1 Origines
- 1.2 Débuts
- 1.3 Les premiers succès
- 1.4 La confirmation

2 Analyse de l'œuvre
- 2.1 Thèmes
- 2.2 Influences

3 Discographie
- 3.1 Albums
- 3.2 Participations

4 Distinctions
5 Notes et références
- 5.1 Sources

6 Lien externe

Biographie [modifier | modifier le code]

Cette section **ne cite pas suffisamment ses sources**. Pour l'améliorer, ajouter en note des références vérifiables ou les modèles {{Référence nécessaire}} ou {{Référence souhaitée}} sur les passages nécessitant une source.

Pierre Lapointe

Pierre Lapointe lors d'un spectacle donné au Café théâtre le Côté-Cour de Jonquière (mars 2011).

Informations générales

Naissance	23 mai 1981 (33 ans) Alma, Québec, Canada
Activité principale	Chanteur, auteur-compositeur-interprète
Genre musical	Chanson française
Instruments	piano, voix
Années actives	Depuis 2000
Labels	Audiogram
Site officiel	www.pierrelapointe.com

- WOE est décrit en 2 saveurs :
  - WOE^{parse} qui utilise un analyseur en dépendance
  - WOE^{pos} qui utilise les POS (comme TEXTRUNNER)

WOE / 1 - distant supervision

Pour chaque article a de WIKIPEDIA, pour chaque paire attribut/valeur (t, v) dans l'infobox de a :

1. matcher dans s — phrase de a — le titre de a (arg_1) et v (arg_2)
   - présence de v :
     - correspondance exacte
     - synonyme est présent (via redirect, in-links) : *UK* vs *United Kingdom*
   - présence de $titre(a)$:
     - correspondance exacte ou synonyme (redirects)
     - partial match (préfixe ou suffixe) : *Pierre* vs *Pierre Lapointe*
     - pronom le plus fréquent : *He* vs *Pierre Lapointe* *sauf si le pronom le plus fréquent est it*
2. filtrer les exemples positifs
   - une seule phrase dans a doit matcher v
   - (arg_1) et v (arg_2) dans la même clause, etc.

Pour plus de 1M d'articles : 301 962 exemples positifs

Pierre Lapointe *(born May 23, 1981 in Alma, Quebec) is a Québécois* ***singer*** *and keyboardist.*

WOE / 2 - extracteur (cas WOE^{parse})

- pour toute phrase positive
  Pierre$_1$ *was*$_2$ *not*$_3$ *born*$_4$ *in*$_5$ ***Montreal***$_5$

```
nsubjpass(4,1)      auxpass(4,2)
neg(4,3)            prep-in(4,6)
```

corePath entre arg$_1$ et arg$_2$: *Pierre* $\xrightarrow{\text{nsubjpass}}$ *born* $\xleftarrow{\text{prep-in}}$ *Montréal*

expandPath ajouts des dép. adjectivales, adverbiales, auxpass, neg. : *was* $\xrightarrow{\text{auxpass}}$ *born* et *not* $\xrightarrow{\text{neg}}$ *born*

pattern généralisation des corePaths : *N* $\xrightarrow{\text{nsubjpass}}$ *V* $\xleftarrow{\text{prep}}$ *N*

- une BDD de patrons avec leur fréquence
  - 15 333 patterns, 185 ayant une fréq. $\geq$100, 1929 $\geq$ 5
  - prob d'un patron est une fonction de la fréquence de ce patron
    p(N $\xrightarrow{\text{nsubjpass}}$ *V* $\xleftarrow{\text{prep}}$ *N) = 0.95*
- les *expandPaths* sont consultés pour extraire les tuples
  {*Pierre*, *was not born in*, *Montreal*}

WOE / 2 - extracteur (cas WOE^{pos})

- même corpus de phrases positives matchées (arg_1, arg_2)
- les mots des *extendPaths* sont les relations (ex : *was not born in*)

  $\{$*Pierre*, *was not born in*, *Montreal*$\}$ *est un exemple positif*
- exemples négatifs sont pris aléatoirement parmi des phrases contenant des NPs dont le *corePath* n'est pas dans la base de patrons
- CRF entraîné sur les POS (non lexicalisé)

très similaire à TEXTRUNNER *(avec* WIKIPEDIA *plutôt que des heuristiques pour sélectionner les exemples positifs)*

rali

WOE / Évaluation

- 300 phrases du WSJ, de WIKIPEDIA et du Web (total : 900)
- annotation manuelle de tous les triplets dans ces phrases
- **tâche :** retrouver ces tuples (précision/rappel)
  - $\text{WOE}^{parse} > \text{WOE}^{pos} > \text{TEXTRUNNER}$
    courbes AUC (Area Under the Curve)
  - nb moy. de tuples par phrase : WOE^{parse} 1.42, WOE^{pos} 1.05, TEXTRUNNER 0.75
  - la performance de WOE^{pos} et de TEXTRUNNER se degrade rapidement sur les phrases longues, celle de WOE^{parse} moins
  - temps moyen pour traiter une phrase : WOE^{parse} 0.679s WOE^{pos} et TEXTRUNNER : 0.022s (30 fois plus rapide)

rali

REVERB

http://openie.cs.washington.edu

- 3 modules
  1. identification d'une relation verbale
  2. identification des arguments (NP) à gauche et à droite
  3. estimation de confiance (maxent)
- Évaluation
  - 500 pages Web retournées par http://random.yahoo.com/bin/ryl soumises à plusieurs extracteurs et dont les sorties sont évaluées manuellement (rappel évalué par l'union des tuples extraits par les différents systèmes comparés)
  - 500M de phrases de ClueWeb
  - les évaluations suggèrent que REVERB est meilleur que TEXTRUNNER et WOE.

REVERB : Extraction des relations

- contrainte syntaxique :

```
V | V P | V W* P
V = verb particle? adv?
W = (noun | adj | adv | pron | det)
P = (prep | particle | inf. marker)
```

  ex : *invented* (V), *located in* (VP), *has atomic weight of* (VW*P)

+ joindre deux relations adjacentes

- contrainte lexicale :
  - pour éviter des relations trop spécifiques
    ex : *is offering only modest greenhouse gas reduction targets at*
  - **idée :** une relation non spécifique doit avoir plusieurs arguments
  - en pratique : première passe pour extraire sur 500M de phrases toutes les relations et leurs arguments, en gardant les relations qui ont au moins 20 paires d'arguments différentes

REVERB : Extraction des arguments

Pour chaque relation r :

- trouver x le NP le plus proche de r à gauche
  - qui n'est pas un pronom relatif, un adverbe WHO, ou `there`
- trouver y le NP le plus proche à droite de r

TOKS	Elle reprend ses études à l' Université McGill et obtient un doctorat en psychologie (1965) .
POS	CLS V DET NC P DET NC NPP CC V DET NC P NC PONCT NC PONCT PONCT
CHNK	B-NP B-VN B-NP I-NP B-PP B-NP I-NP I-NP B-COORD B-VN B-NP I-NP B-PP B-NP O B-NP O O

REVERB : Extraction des arguments

Pour chaque relation r :

- trouver x le NP le plus proche de r à gauche
  - qui n'est pas un pronom relatif, un adverbe WHO, ou `there`
- trouver y le NP le plus proche à droite de r

TOKS	Elle reprend ses études à l' Université McGill et obtient un doctorat en psychologie (1965) .
POS	CLS V DET NC P DET NC NPP CC V DET NC P NC PONCT NC PONCT PONCT
CHNK	B-NP B-VN B-NP I-NP B-PP B-NP I-NP I-NP B-COORD B-VN B-NP I-NP B-PP B-NP O B-NP O O
EXTR	Elle == reprend == ses études

REVERB : Extraction des arguments

Pour chaque relation r :

- trouver x le NP le plus proche de r à gauche
  - qui n'est pas un pronom relatif, un adverbe WHO, ou `there`
- trouver y le NP le plus proche à droite de r

TOKS	Elle reprend ses études à l' Université McGill et obtient un doctorat en psychologie (1965) .
POS	CLS V DET NC P DET NC NPP CC V DET NC P NC PONCT NC PONCT PONCT
CHNK	B-NP B-VN B-NP I-NP B-PP B-NP I-NP I-NP B-COORD B-VN B-NP I-NP B-PP B-NP O B-NP O O
EXTR	Elle == reprend == ses études
EXTR	ses études == obtient == un doctorat

REVERB : Estimateur de confiance

- **in :** (x, r, y) **out** : $p(\text{ok})$
- entraîné sur les évaluations manuelles (correcte ou pas) des extractions réalisées sur un corpus de 1000 phrases
- 19 features qui ne dépendent pas d'une relation en particulier :

1.16	(x, r, y) couvre tous les mots de s
0.50	la dernière prép dans r est *for*
0.49	la dernière prép dans r est *on*
0.46	la dernière prép dans r est *of*
0.43	length(s) $\leq$ 10 mots
-0.93	conjonction de coord. à gauche de r dans s
	$\vdots$

OLLIE *[Mausam et al., 2012]*

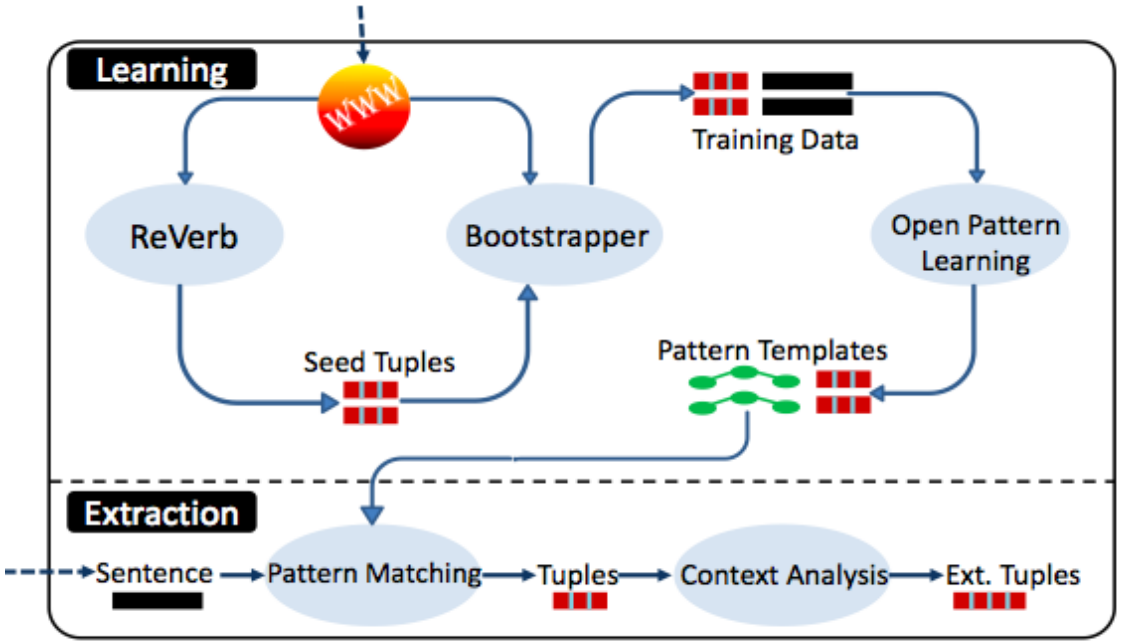


Fig 2 dans l'article

- apprend des patrons syntaxiques à partir de tuples extraits par REVERB
- reconnaît des relations nominales : *Microsoft co-founder Bill Gates said. . .* $\Rightarrow$ {Bill Gates, is cofounder of, Microsoft}

OLLIE / Learning

- ∼110k tuples les plus fiables extraits par REVERB depuis CLUEWEB :
  - freq≥2
  - args = noms propres
  - confiance élevée (selon REVERB)
- 18M des phrases de CLUEWEB contiennent ces tuples
- 4M une fois filtrées (mots de têtes des args reliés par un chemin de dépendances de long. au plus 4)
- **hyp** : ces phrases expriment des relations pertinentes
  vrai à 90% selon une éval. manuelle sur 100 phrases

OLLIE / Learning

Extraction Template	Open Pattern
1. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓{prep_*}↓ {arg2}
2. (arg1; {rel}; arg2)	{arg1} ↑nsubj↑ {rel:postag=VBD} ↓dobj↓ {arg2}
3. (arg1; be {rel} by; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓agent↓ {arg2}
4. (arg1; be {rel} of; arg2)	{rel:postag=NN;**type=Person**} ↑nn↑ {arg1} ↓nn↓ {arg2}
5. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {slot:postag=VBN;**lex** ∈**announce**|**name**|**choose...**} ↓dobj↓ {rel:postag=NN} ↓{prep_*}↓ {arg2}

Tab 3 dans l'article

À partir de l'analyse en dépendance des 4M de phrases retenues, OLLIE extrait des patrons syntaxiques (droite) qui expriment des relations (gauche) :

1. des patrons purement syntaxiques (ex : 1 à 3)
   - **fiables :** sans slot nodes, *rel* entre arg_1 et arg_2, etc.
   - généralisation aggressive (*on*, *on*, *for*, . . . = prep)

OLLIE / Learning

Extraction Template	Open Pattern
1. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓{prep_*}↓ {arg2}
2. (arg1; {rel}; arg2)	{arg1} ↑nsubj↑ {rel:postag=VBD} ↓dobj↓ {arg2}
3. (arg1; be {rel} by; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓agent↓ {arg2}
4. (arg1; be {rel} of; arg2)	{rel:postag=NN;**type=Person**} ↑nn↑ {arg1} ↓nn↓ {arg2}
5. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {slot:postag=VBN;**lex** ∈**announce**|**name**|**choose...**} ↓dobj↓ {rel:postag=NN} ↓{prep_*}↓ {arg2}

Tab 3 dans l'article

À partir de l'analyse en dépendance des 4M de phrases retenues, OLLIE extrait des patrons syntaxiques (droite) qui expriment des relations (gauche) :

1. des patrons purement syntaxiques (ex : 1 à 3)
   - **fiables :** sans slot nodes, *rel* entre arg_1 et arg_2, etc.
   - généralisation aggressive (*on*, *on*, *for*, . . . = prep)
2. des patrons lexicalisés (ex : 5)
   - avec slot nodes

OLLIE / Learning

Extraction Template	Open Pattern
1. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓{prep_*}↓ {arg2}
2. (arg1; {rel}; arg2)	{arg1} ↑nsubj↑ {rel:postag=VBD} ↓dobj↓ {arg2}
3. (arg1; be {rel} by; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓agent↓ {arg2}
4. (arg1; be {rel} of; arg2)	{rel:postag=NN;**type=Person**} ↑nn↑ {arg1} ↓nn↓ {arg2}
5. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {slot:postag=VBN;**lex** ∈**announce**|**name**|**choose...**} ↓dobj↓ {rel:postag=NN} ↓{prep_*}↓ {arg2}

Tab 3 dans l'article

À partir de l'analyse en dépendance des 4M de phrases retenues, OLLIE extrait des patrons syntaxiques (droite) qui expriment des relations (gauche) :

1. des patrons purement syntaxiques (ex : 1 à 3)
   - **fiables :** sans slot nodes, *rel* entre arg_1 et arg_2, etc.
   - généralisation aggressive (*on*, *on*, *for*, . . . = prep)
2. des patrons lexicalisés (ex : 5)
   - avec slot nodes
3. des patrons typés (ex : 4)
   - généralisation des patrons lexicalisés (grâce à WORDNET)

OLLIE / Extraction

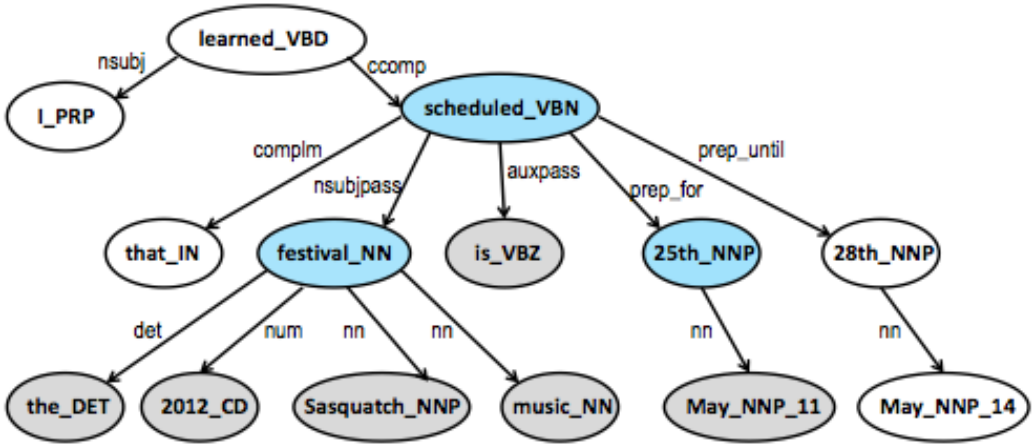


Fig 4 dans l'article

I learned that the 2012 Sasq. music festival is scheduled for May 25th until May 28th

- template : (arg ; be {rel} {prep} ; arg2)
- patron : {arg1} $\xrightarrow{nsubjpass}$ {rel :postag=VBN} $\xleftarrow{prep}$ {arg2}

1. match arg1=*festival* arg2=*25th* et rel=*scheduled*
   {*festival*, *be scheduled for*, *25th*}

OLLIE / Extraction

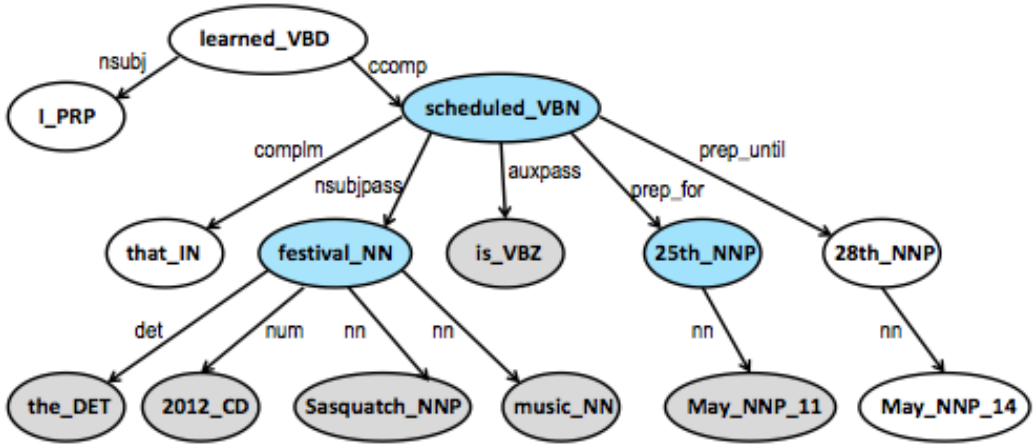


Fig 4 dans l'article

I learned that the 2012 Sasq. music festival is scheduled for May 25th until May 28th

- template : (arg ; be {rel} {prep} ; arg2)
- patron : {arg1} $\xrightarrow{nsubjpass}$ {rel :postag=VBN} $\xleftarrow{prep}$ {arg2}

1. match arg1=*festival* arg2=*25th* et rel=*scheduled*
   {*festival*, *be scheduled for*, *25th*}
2. étendre les nœuds args et rel (voir l'article pour les détails)
   {*the Sasquatch music festival*, *be scheduled for*, *May 25th*}

OLLIE / Extraction

- OLLIE détecte 2 cas d'extractions non factuelles :

  attribution celui qui *dit*/ *croit*/ *pense*/etc.
  Early astronomers believed that the earth is the center of the universe

  R : {the earth, be the center of, the universe}
  O : ({the earth, be the center of, the universe}
  AttributedTo : believe ; Early astronomers)

  modificateur extraction conditionnellement vraie
  If he wins five states, Romney will be elected President

  R : {Romney, will be elected, President}
  O : ({Romney, will be elected, President}
  ClausalModifier if ; he wins five key states)

- les autres cas sont laissés à un estimateur de confiance (entraîné sur 1000 phrases de domaines testés. . .)

OLLIE / Évalution

- sur 300 phrases de WIKIPEDIA, de NEWS et de livres de biologie
- annotation manuelle des tuples extraits par WOE^{parse}, REVERB et OLLIE(1945 tuples)
- ref=tuples annotés corrects par deux annotateurs (accord : 0.96)
  biais en faveur de la précision
  - OLLIE trouve en gros 4 fois plus de tuples corrects que les autres extracteurs à une précision de 75%
  - usage du parseur explique une bonne partie de ce gain (% à REVERB)
  - relations non lexicalisées entre les arguments
    After winning the Superbowl, the Saints are now the top dogsbof the NFL $\Rightarrow$ {the Saints, win, the Superbowl}
  - OLLIE extrait des relations nominales, WOE^{parse} non :
    Obama, the president of the US $\Rightarrow$
    {Obama, be the president of, the US}

[Mesquita et al., 2013]

- une évaluation de 8 extracteurs (3 familles + EXAMPLAR) :
  - POS tagger : REVERB, SONEX
  - analyseur en dépendance : OLLIE, TREEKERNEL, PATTY
  - étiqueteur sémantique (SRL) : LUND, SWIRL

  - 500 phrases du NYT (461), 500 du Web (150), 100 du PTB (51)
  - annotation manuelle de toutes les relations (binaires ou pas)
  - entités nommées remplacées par *Europe* et *Asia*
- un extracteur — EXAMPLAR — plus performant :
  - système à base de règles résultant d'une analyse linguistique de relations n-aires ($n \geq 2$)
  - comme un SRL, mais basé uniquement sur les dépendances syntaxiques

[Mesquita et al., 2013]

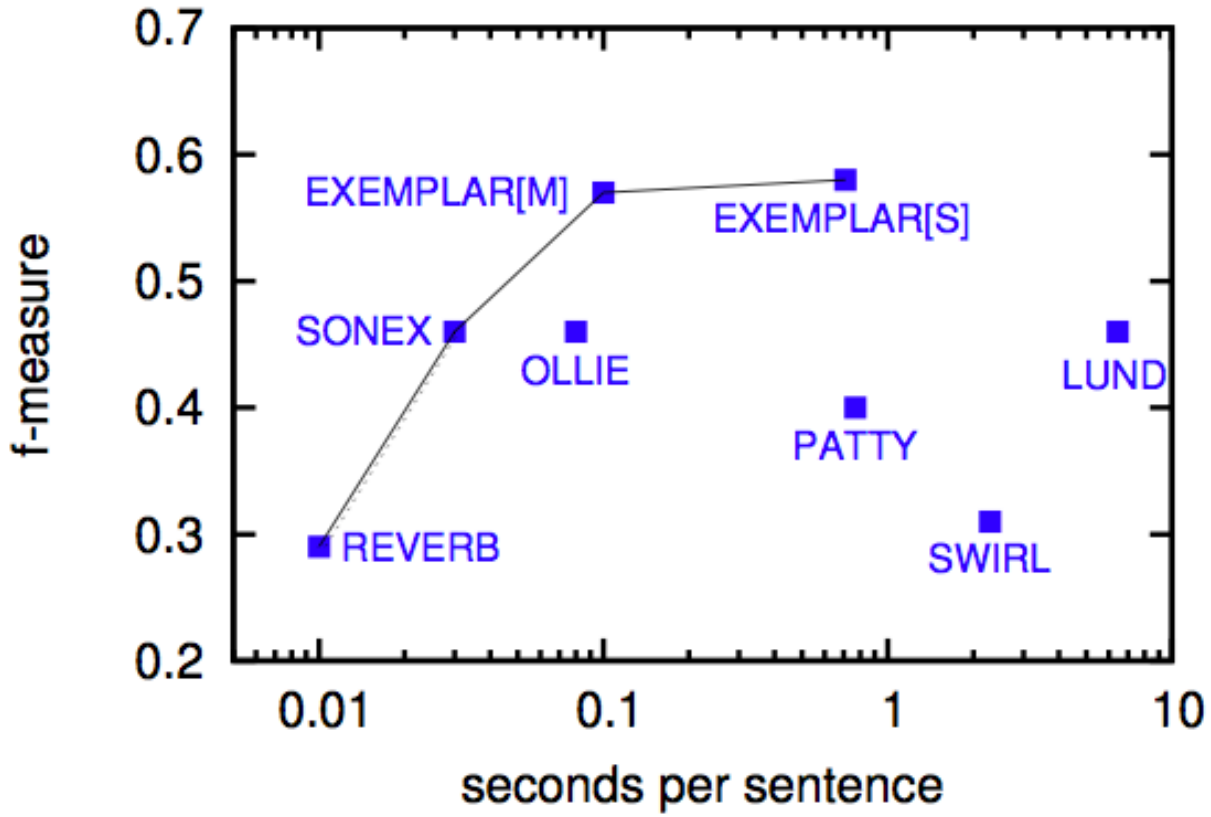


Figure 1: Average f-measure vs average time for NYT-500, WEB-500 and PEN-100.

[Mesquita et al., 2013]

subject:	"NFL"
relation:	"approve new stadium"
of_object:	"Falcons"
in_object:	"Atlanta"

Figure 3: A relation instance extracted by EXEMPLAR for the sentence "NFL approves Falcons' new stadium in Atlanta".

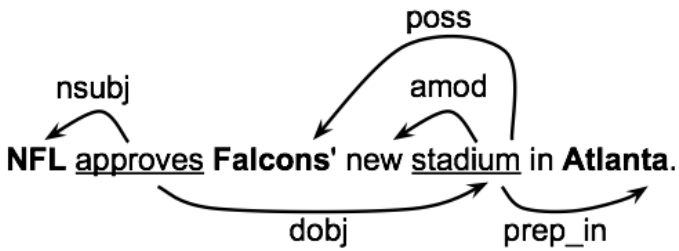


Figure 4: A input sentence after pre-processing. Entities are in bold, triggers are underline and arrows represent dependencies.

- EXAMPLAR identifie :
  1. triggers : mots dont la présence indique une des 3 relations
     - verbe *Hingins **beat** Steffi Graf in the Italian Open two weeks ago*
     - copule+nom *Kimball **was a** Fullbright **scholar** at uni. of New York*
     - verbe+nom *D-League will **move its offices** from X to Y*
  2. arguments (les acteurs potentiels de la relation)
  3. roles (le rôle de chaque argument)
- + filtrages

Plan

Extracteurs

Structuration

Applications

Analyse

Datasets

PATTY [Nakashole et al., 2012]

- Extracteur textuel : très similaire à OLLIE
  - identification d'entités nommées en utilisant une ressource dédiée (YAGO2 dans leur cas)
  - extraction d'un patron textuel à partir d'un arbre de dépendance (même poutine : shortest-path + ajout de dépendants (adv, etc.))
- Généralisation à l'aide des patrons Syntaxic-Ontological-Lexical
  - séquence de POS, de $\star$ et de types (eg : `<songwriter>`)
  - ex : <person>'s [adj] voice * <song> matche :
    Amy Winehouse's soft voice in 'Rehab'
    Elvis Presley's solid voice in this song 'All shook up'
  - signature d'un patron est la paire de types (person×song)
  - ensemble support : les paires d'instances qui marchent le patron ({(Amy,Rehab), (Elvis,AllShookUp)})

PATTY [Nakashole et al., 2012]

- passage des instances aux patrons SOL :
  - découpage du patron textuel en n-grams, les n-grams ($n = 3$) fréquents sont retenus, les autres deviennent $\star$
  - généralisation soit en introduisant des $\star$, en remplaçant les mots par leur POS, ou en généralisant les types
  - généralisation rejetée si elle domine des patrons qui ont des supports disjoints (voir détails)
- structuration selon :
  - un patron A est plus général qu' (ou domine) un patron B si le support set de B inclus dans celui de A
  - A et B sont des patrons synonymes si A est plus général que B et que B est plus général que A (même support)
- **output :** DAG de synsets de patrons dont les arcs indiquent les relations de dominance

PATTY / Évaluation

- www.mpi-inf.mpg.de/yago-naga/patty/

  NYT 1.8M d'articles du New-York Time (1987 – 2007)
  WKP 3.8M d'articles de WIKIPEDIA (dump de juin 2011)

- Évaluation sur 100 patrons les mieux *supportés* :

  qualité NYT : 72%, WKP : 85%
  dominance NYT : 86%, WKP : 83%
  $< person >$ *nominated for* $< award >$ $\Leftarrow$
  $< person >$ *winner of* $< award >$

- Évaluation de la dominance impliquant 100 patrons choisis aléatoirement : NYT : 68%, WPT : 75%
- rappel sur les relations impliquant le type musician dans 5 articles de WIKIPEDIA (*Amy Winehouse*, *B. Dylan*, *N. Young*, *J. Coltrane* et *N. Simone*) :
  - 163 relations identifiées manuellement et communément par 2 annotateurs
  - 126 relations identifiées par PATTY, 31 dans YAGO2, 39 dans DBPEDIA, 69 dans FREEBASE et 13 dans NELL.

PATTY / Évaluation

- aptitude à paraphraser une relation

Relation	Paraphrases	Precision	Sample Paraphrases
DBPedia/artist	83	0.96±0.03	[adj] studio album of, [det] song by ...
DBPedia/associatedBand	386	0.74±0.11	joined band along, plays in ...
DBPedia/doctoralAdvisor	36	0.558±0.15	[det] student of, under * supervision ...
DBPedia/recordLabel	113	0.86±0.09	[adj] artist signed to, [adj] record label ...
DBPedia/riverMouth	31	0.83±0.12	drains into, [adj] tributary of ...
DBPedia/team	1,108	0.91±0.07	be * traded to, [prp] debut for ...
YAGO/actedIn	330	0.88±0.08	starred in * film, [adj] role for ...
YAGO/created	466	0.79±0.10	founded, 's book ...
YAGO/isLeaderOf	40	0.53±0.14	elected by, governor of ...
YAGO/holdsPoliticalPosition	72	0.73±0.10	[prp] tenure as, oath as ...

Table 6: Sample Results for Relation Paraphrasing

rali

WEBRE *[Min et al., 2012]*

in triplets extraits par REVERB de CLUEWEB

pré-calcul de sources de connaissances

DS Distributional Similarity
PS Pattern Similarity

1. patrons lexicaux
2. patrons HTML

WEBRE *[Min et al., 2012]*

in triplets extraits par REVERB de CLUEWEB

pré-calcul de sources de connaissances

DS Distributional Similarity

- des termes qui partagent des contextes similaires sont similaires *[Harris, 1985]*
- contexte=fenêtres de taille 4, association=PMI similarité=Jaccard

PS Pattern Similarity

1. patrons lexicaux
2. patrons HTML

WEBRE *[Min et al., 2012]*

in triplets extraits par REVERB de CLUEWEB

pré-calcul de sources de connaissances

DS Distributional Similarity

PS Pattern Similarity

1. patrons lexicaux
   - 2 termes similaires = partagent bcp de patrons lexicaux
   - ex : *(such as | including) T{,T}⋆(, and |.)*
2. patrons HTML
   - souvent listés dans des structures HTML similaires
   - ex : tables, menus déroulants, . . .

rali

WEBRE *[Min et al., 2012]*

in triplets extraits par REVERB de CLUEWEB

pré-calcul de sources de connaissances

DS Distributional Similarity

PS Pattern Similarity

1. patrons lexicaux
2. patrons HTML

29.1M noeuds, 1.12 billion d'arcs

HG Hypernymy Graph

- similarité = partage d'hypernyme
- via patrons : *NP* {,} *such as* {*NP*,}⋆ {*and*|*or*} *NP*

WEBRE *[Min et al., 2012]*

in triplets extraits par REVERB de CLUEWEB

pré-calcul de sources de connaissances

- **DS** Distributional Similarity
- **PS** Pattern Similarity
  1. patrons lexicaux
  2. patrons HTML

  29.1M noeuds, 1.12 billion d'arcs
- **HG** Hypernymy Graph

  8.2M noeuds, 42.4M d'arcs hypo→hyper
- **RS** Relation-Phrase Similarity
  - hypothèse distributionnelle (ici partage d'arguments)

out collections d'items de type A et de type B

- **A** {{Obama, Clinton,...}, win in, {PA,CA,...}}
- **B** {{Obama,...}, {win in, take,...}, {PA,CA,...}}

WEBRE *[Min et al., 2012]*

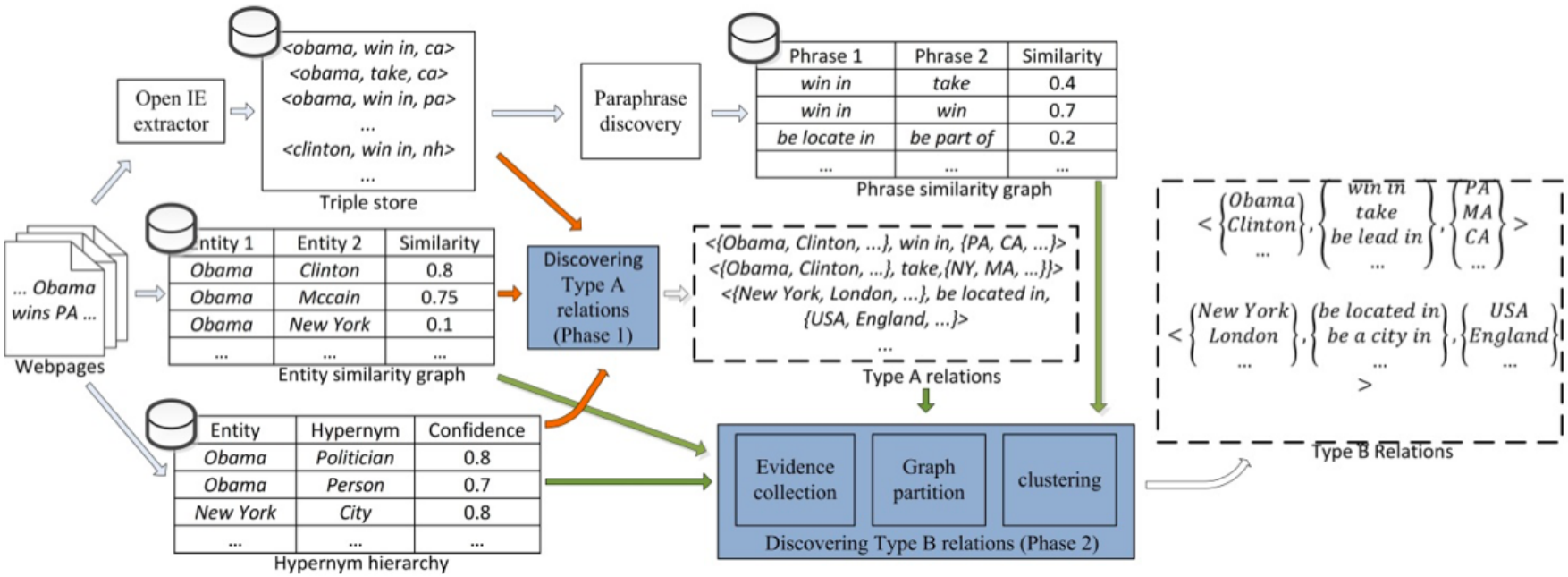


Figure 1. Overview of the WEBRE algorithm (Illustrated with examples sampled from experiment results). The tables and rectangles with a database sign show knowledge sources, shaded rectangles show the 2 phases, and the dotted shapes show the system output, a set of Type A relations and a set of Type B relations. The orange arrows denote resources used in phase 1 and the green arrows show the resources used in phase 2.

WEBRE *[Min et al., 2012]*

- algo de regroupement efficace

  type A

  1. pour une relation verbale, clusteriser les arg_1
  2. pour chaque regroupement de arg_1, réunir les arg_2 apparaissant dans les triplets avec ces arg_1
  3. clusteriser les ensembles d'arg_2 ainsi obtenus

  type B

  1. partition des relations de type A
  2. clustering the chaque partition
     - en utilisant les ressources
     - *détails (un peu) absconds*

- temps sur une seule machine (one CPU core) :
  22 heures pour type A, 4h pour type B

WEBRE *[Min et al., 2012]*

- 14.7M de triplets distincts extraits par REVERB depuis CLUEWEB (filtres appliqués), 1.3M relations verbales, 3.3M entités (args)
- 0.2M relations de type A, 84 000 de type B

Argument 1	Relation phrase	Argument 2
marijuana, caffeine, nicotine...	*result in, be risk factor for, be major cause of...*	*insomnia, emphysema, breast cancer,...*
C# 2.0, php5, java, c++, ...	*allow the use of, also use, introduce the concept of...*	*destructors, interfaces, template,...*
clinton, obama, mccain, ...	*win, win in, take, be lead in,...*	*ca, dc, fl, nh, pa, va, ga, il, nc,...*

Table 3. Sample Type B relations extracted.

[Balasubramanian et al., 2013]

- Extraction de schémas d'événements à partir de textes

Actor	Rel	Actor
A1:<person>	failed	A2:test
A1:<person>	was suspended for	A3:<time period>
A1:<person>	used	A4:<substance, drug>
A1:<person>	was suspended for	A5:<game, activity>
A1:<person>	was in	A6:<location>
A1:<person>	was suspended by	A7:<org, person>
Actor Instances:		
A1: {Murray, Morgan, Governor Bush, Martin, Nelson}		
A2: {test}		
A3: {season, year, week, month, night}		
A4: {cocaine, drug, gasoline, vodka, sedative}		
A5: {violation, game, abuse, misfeasance, riding}		
A6: {desert, Simsbury, Albany, Damascus, Akron}		
A7: {Fitch, NBA, Bud Selig, NFL, Gov Jeb Bush}		

Table 1: An event schema produced by our system, represented as a set of $(Actor, Rel, Actor)$ triples, and a set of instances for each actor $A1$, $A2$, etc. For clarity we show unstemmed verbs.

- *afin de renouer avec l'approche plus traditionnelle à l'extraction d'information*

Plan

Extracteurs

Structuration

Applications

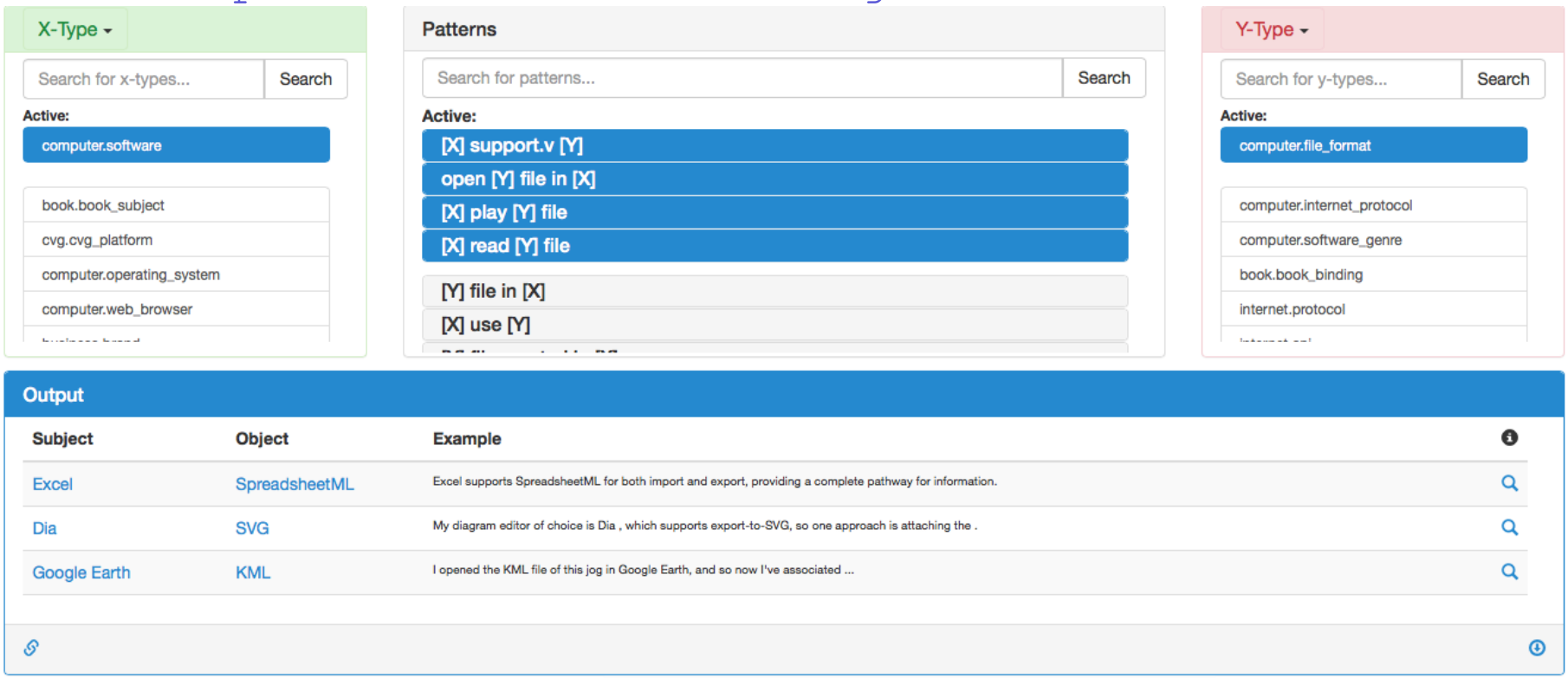
Analyse

Datasets

Exploration of large collections of texts

[Akbik et al., 2014]

http://lucene.textmining.tu-berlin.de/



Open-domain QA *[Fader et al., 2014]*

- QA $\equiv$ set of operators that map a question in natural language into a query to a tuple-store
  - 4 $\neq$ tuple-stores (FREEBASE and 3 automatically extracted ones)
- a scoring function is learnt from question-answer pairs (structured perceptron)
- answering $\equiv$ beam-search

Open-domain QA *[Fader et al., 2014]*

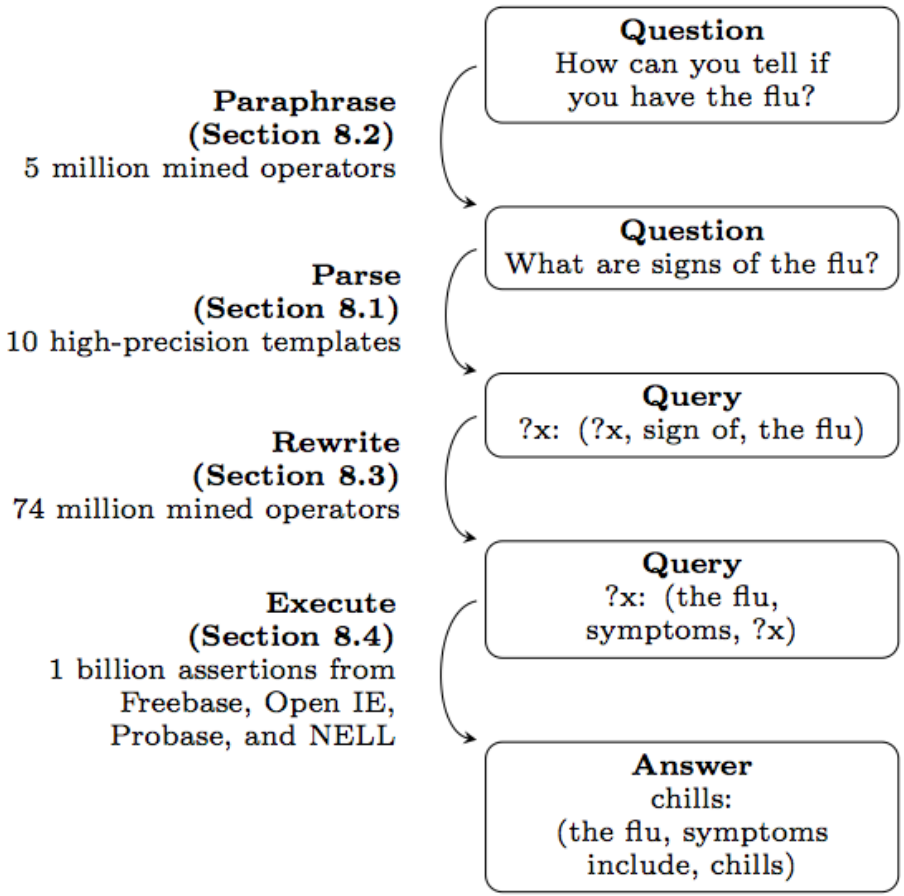


Figure 1: OQA automatically mines millions of operators (left) from unlabeled data, then learns to compose them to answer questions (right) using evidence from multiple knowledge bases.

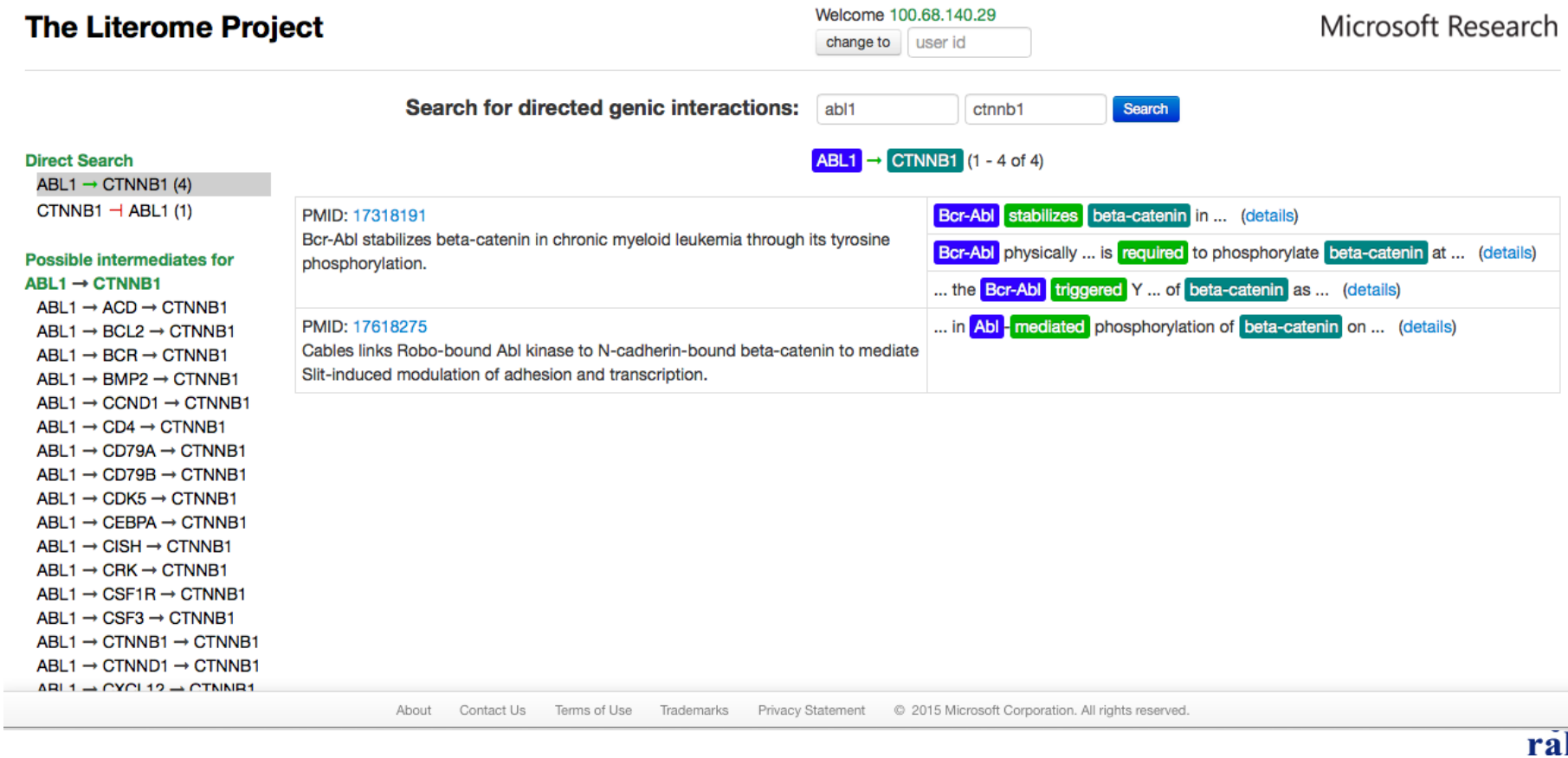
Open-domain QA *[Fader et al., 2014]*

Input	What are some examples of building maintenance jobs ?
Parse	`?x: (?x, example of, building maintenance jobs)`
Rewrite	`?x: (?x, is-a, building maintenance job)`
Execute	{changing light bulb, is-a, small building maintenance job}

Input	What animal represents California ?
Paraphrase	What are California's symbols ?
Parse	`?x: (california,symbols, ?x)`
Execute	{California Water Service, Trading symbol, CWT}

Literome project *[Poon et al., 2014]*

- extraction of genomic knowledge from PubMed articles
- available in the cloud for browsing, searching and reasoning



Plan

Data Analysis

- ran OLLIE and REVERB on the WIKIPEDIA article : Quebec

Article Talk Read View source View history Search

Quebec

From Wikipedia, the free encyclopedia

Coordinates: 53°N 07

This article is about the Canadian province. For the province's capital city, see Quebec City. For other uses, see Quebec (disambiguation).

Quebec (/kwɨˈbɛk/ or /keɪˈbɛk/; French: *Québec* [kebɛk] (listen))[7] is a province in east-central Canada.[8][9] It is the only Canadian province that has a predominantly French-speaking population, and the only one to have French as its sole provincial official language.

Quebec is Canada's largest province by area and its second-largest administrative division; only the territory of Nunavut is larger. It is bordered to the west by the province of Ontario, James Bay, and Hudson Bay; to the north by Hudson Strait and Ungava Bay; to the east by the Gulf of Saint Lawrence and the province of Newfoundland and Labrador; it is bordered on the south by the province of New Brunswick and the U.S. states of Maine, New Hampshire, Vermont, and New York. It also shares maritime borders with Nunavut, Prince Edward Island, and Nova Scotia.

Quebec
***Québec* (French)**

Flag Coat of arms

- on 171 sentences : OLLIE found 445 tuples, REVERB found 217
- we inspected the 100 most likely ones

Many `the X` like arguments

- roughly 50% of the tuples
- mostly useless

Jardin zoologique du Québec , reopened in 2002 after two years of restorations but closed in 2006 after a political decision . It featured 750 specimens of 300 different species of animals . **The zoo specialized in winged fauna and garden themes** , but also presented several species of mammals .

{The zoo, specialized in, winged fauna and garden themes}

rali

Many `the X` like arguments

- roughly 50% of the tuples
- mostly useless

Jardin zoologique du Québec , reopened in 2002 after two years of restorations but closed in 2006 after a political decision . It featured 750 specimens of 300 different species of animals . **The zoo specialized in winged fauna and garden themes** , but also presented several species of mammals .

{The zoo, specialized in, winged fauna and garden themes}

rali

Many `the X` like arguments

- roughly 50% of the tuples
- mostly useless

Porte St-Louis and Porte St-Jean are the main gates through the walls from the modern section of downtown ; the Kent Gate was a gift to the province from Queen Victoria and **the foundation stone** **was laid by the Queen 's daughter** , Princess Louise , Marchioness of Lorne , on June 11 , 1879 .

{the foundation stone, was laid by, the Queen 's daughter}

rali

Many `the X` like arguments

- roughly 50% of the tuples
- mostly useless

Porte St-Louis and Porte St-Jean are the main gates through the walls from the modern section of downtown ; the Kent Gate was a gift to the province from Queen Victoria and **the foundation stone was laid by the Queen 's daughter** , Princess Louise , Marchioness of Lorne , on June 11 , 1879 .

{the foundation stone, was laid by, the Queen 's daughter}

Other coreference issues

Quebec City 's skyline is dominated by the massive Château Frontenac Hotel , perched on top of Cap-Diamant . **It was designed by architect Bruce Price** , as one of a series of " château " style hotels built for the Canadian Pacific Railway company .

{It, was designed by, architect Bruce Price}

Other coreference issues

Quebec City 's skyline is dominated by the massive Château Frontenac Hotel , perched on top of Cap-Diamant . **It was designed by architect Bruce Price** , as one of a series of " château " style hotels built for the Canadian Pacific Railway company .

{It, was designed by, architect Bruce Price}

Other coreference issues

During World War II , two conferences were held in Quebec City . The First Quebec Conference was held in 1943 with Franklin Delano Roosevelt – the United States ' president at the time – , Winston Churchill – the United Kingdom 's prime minister – , William Lyon Mackenzie King – Canada 's prime minister – and T.V. Soong – China 's minister of foreign affairs . The Second Quebec Conference was held in 1944 , and was attended by Churchill and Roosevelt . They took place in the buildings of the Citadelle and of nearby Château Frontenac . **A large part of the D-Day landing plans were made during those meetings** .

{ A large part of the D-Day landing plans; were made during; those meetings }

Other coreference issues

During World War II , two conferences were held in Quebec City . The First Quebec Conference was held in 1943 with Franklin Delano Roosevelt – the United States ' president at the time – , Winston Churchill – the United Kingdom 's prime minister – , William Lyon Mackenzie King – Canada 's prime minister – and T.V. Soong – China 's minister of foreign affairs . The Second Quebec Conference was held in 1944 , and was attended by Churchill and Roosevelt . They took place in the buildings of the Citadelle and of nearby Château Frontenac . **A large part of the D-Day landing plans were made during those meetings** .

{ A large part of the D-Day landing plans; were made during; those meetings }

Many other problems. . .

- Many tuples are uninformative (often by lack of context)

  {The town, distinguished, itself}
  {One-quarter of the people, were members of, religious}
  {The team, has, size league titles}

- Syntax is hard

The English-speaking community peaked in relative terms during the 1860s , when 40 % of Quebec City 's residents were Anglophone .

{The English-speaking community, peaked in, relative terms}

rali

Yet it works !

- provided you mine large quantities of texts
  millions of sentences
- and apply agressive filtering
  high confidence scores, seen at least n times, etc.

$\{$Quebec, was founded by, Samuel de Champlain$\}$
$\{$Quebec City, is located on, the north bank of the Saint Lawrence River$\}$
$\{$Quebec City, is an important hub in, the province 's autoroute system$\}$
$\{$Quebec City, is located in, the Saint Lawrence River valley$\}$
$\{$Quebec City, was struck by, the 1925 Charlevoix-Kamouraska earthquake$\}$

rali

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
   20M tuples once filtered

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories
3. computed the **relational profile** of each category : most discriminative relations involving instances of a category

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories
3. computed the **relational profile** of each category : most discriminative relations involving instances of a category
   Philosophe $\rightarrow$ *affirmer, appeler, considérer, décrire, défendre, développer, fonder, publier, reprendre, écrire, etc.*

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories
3. computed the **relational profile** of each category : most discriminative relations involving instances of a category
4. selected the most representative instances of each category thanks to their profile

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories
3. computed the **relational profile** of each category : most discriminative relations involving instances of a category
4. selected the most representative instances of each category thanks to their profile
   Actor $\rightarrow$ *David Arquette, Jeremy Bulloch, Fernand Gravey, etc.*

Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories
3. computed the **relational profile** of each category : most discriminative relations involving instances of a category
4. selected the most representative instances of each category thanks to their profile
5. clustered (K-means) the categories according to their profile

Interlude : DEFACTO

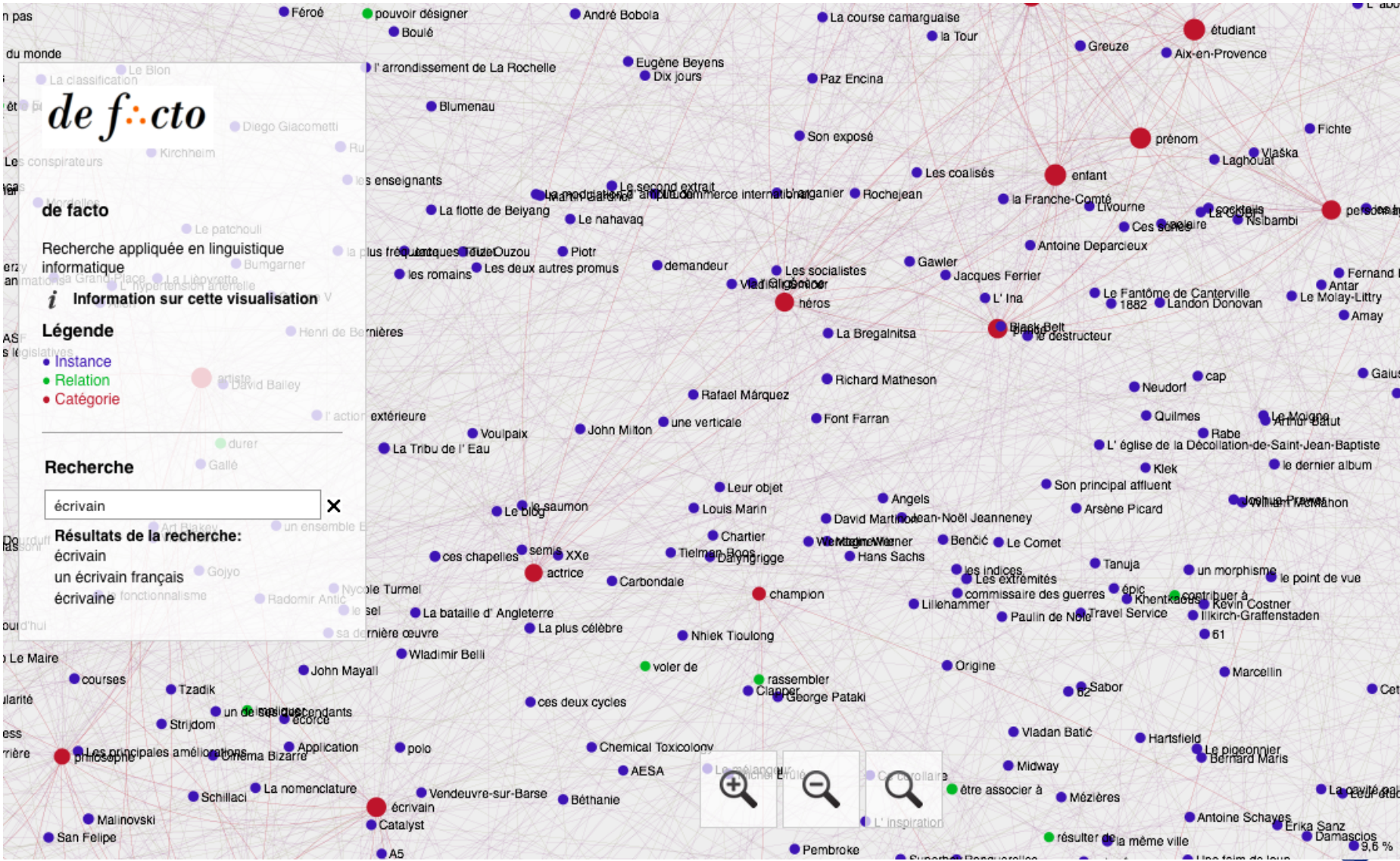
1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories
3. computed the **relational profile** of each category : most discriminative relations involving instances of a category
4. selected the most representative instances of each category thanks to their profile
5. clustered (K-means) the categories according to their profile
   Écrivain$\rightarrow$ *artiste, auteur, créateur, dessinateur, historien, magazine, philosophe, élève*

rali

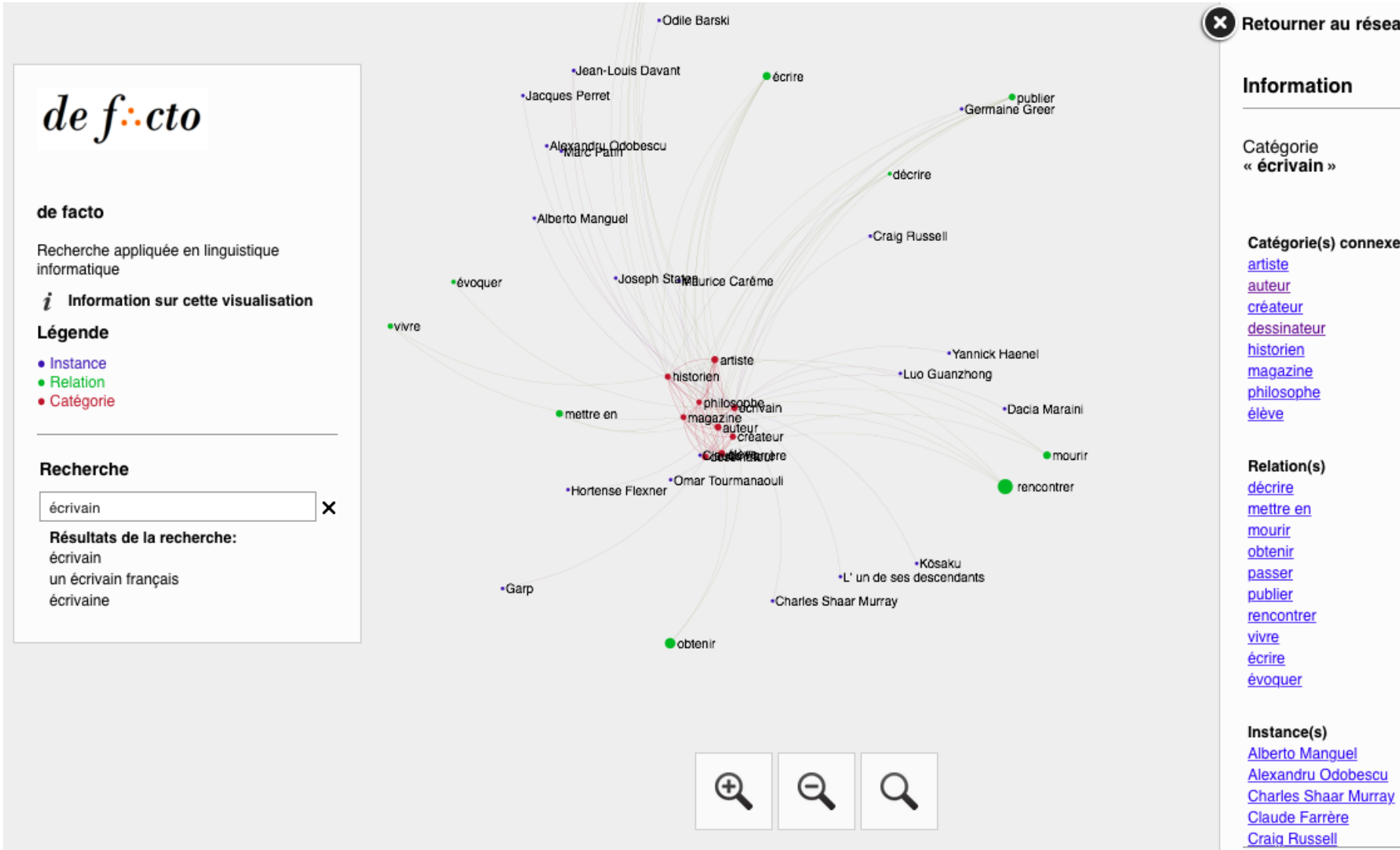
Interlude : DEFACTO

1. extracted tuples from 31M sentences of WIKIPEDIA-FR with an in-house extractor (we adapted REVERB to French)
2. most fréquent arg_2s of relation *être/is/* elected categories
3. computed the **relational profile** of each category : most discriminative relations involving instances of a category
4. selected the most representative instances of each category thanks to their profile
5. clustered (K-means) the categories according to their profile
   Écrivain$\rightarrow$ *artiste, auteur, créateur, dessinateur, historien, magazine, philosophe, élève*
6. generated a (huge) graph with all this, navigable through a plugin (`sigma.js`) we modified for the purpose

DEFACTO


de f:.cto
de facto
Recherche appliquée en linguistique informatique
Information sur cette visualisation
Légende
Instance
Relation
Catégorie
Recherche
écrivain
Résultats de la recherche:
écrivain
un écrivain français
écrivaine
Féroé
pouvoir désigner
André Bobola
La course camargualse
la Tour
étudiant
Greuze
Aix-en-Provence
Boulé
l' arrondissement de La Rochelle
Eugène Beyens
Dix jours
Paz Encina
Blumenau
Son exposé
prénom
Fichte
Vlaška
Laghouat
Les coalisés
enfant
Rochejean
la Franche-Comté
Livourne
Ns bambi
Le second extrait
La flotte de Belyang
Le nahavaq
Antoine Deparcieux
Piotr
Les deux autres promus
les romains
demandeur
Les socialistes
Gawler
Jacques Ferrier
héros
L' Ina
Le Fantôme de Canterville
1882
Landon Donovan
Le Molay-Littry
Antar
Amay
La Bregalnitsa
Richard Matheson
Rafael Márquez
Neudorf
cap
Quilmes
Rabe
Font Farran
une verticale
John Milton
Voulpaix
La Tribu de l' Eau
L' église de la Décollation-de-Saint-Jean-Baptiste
le dernier album
Klek
Son principal affluent
Leur objet
Angels
Arsène Picard
Louis Marin
Chartier
Benčić
Le Comet
Hans Sachs
Tanuja
un morphisme
le point de vue
Les extrémités
commissaire des guerres
Lillehammer
Paulin de Noé
Travel Service
Kevin Costner
Ilkirch-Graffenstaden
61
ces chapelles
semis
XXe
actrice
Carbondale
champion
La bataille d' Angleterre
La plus célèbre
Nhiek Tioulong
Origine
Marcellin
Wladimir Belli
voler de
rassembler
George Pataki
ces deux cycles
courses
Tzadik
John Mayall
Strjdom
Application
polo
Chemical Toxicolony
AESA
Vladan Batić
Hartsfield
Le pigeonnier
Bernard Maris
Midway
être associer à
Mézières
Schillaci
La nomenclature
Vendeuvre-sur-Barse
Béthanie
écrivain
Catalyst
Malinovski
San Felipe
A5
Pembroke
Antoine Schaves
Erika Sanz
Damascios
résulter de
la même ville

DEFACTO


de f::cto
de facto
Recherche appliquée en linguistique informatique
Information sur cette visualisation
Légende
Instance
Relation
Catégorie
Recherche
écrivain
Résultats de la recherche:
écrivain
un écrivain français
écrivaine
Odile Barski
Jean-Louis Davant
écrire
Jacques Perret
publier
Germaine Greer
Alexandru Odobescu
décrire
Alberto Manguel
Craig Russell
évoquer
Joseph Staline
Maurice Carême
vivre
artiste
historien
Yannick Haenel
Luo Guanzhong
philosophe
écrivain
mettre en
magazine
auteur
créateur
Dacia Maraini
mourir
Omar Tourmanaouli
Hortense Flexner
rencontrer
Kōsaku
L' un de ses descendants
Garp
Charles Shaar Murray
obtenir
Retourner au réseau
Information
Catégorie
« écrivain »
Catégorie(s) connexe
artiste
auteur
créateur
dessinateur
historien
magazine
philosophe
élève
Relation(s)
décrire
mettre en
mourir
obtenir
passer
publier
rencontrer
vivre
écrire
évoquer
Instance(s)
Alberto Manguel
Alexandru Odobescu
Charles Shaar Murray
Claude Farrère
Craig Russell
rali

What we have learnt so far

- extractors are (far) suboptimal
- structuring information is hard
- not much awareness to domain (open IE after all !)

Plan

Crédit

- la plupart de ces ressources (et d'autres) sont listées ici : http://www.cs.cmu.edu/~mfaruqui/suite.html

WordSimilarity-353

- http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/
  - **set1** 153 paires de mots : 13 jugements de relatedness
  - **set2** 200 paires de mots : 16 jugements

```
% more set1.csv | sort -t\, -k3,3n+
tiger,tiger,10.00,10,10,10,10,10,10,10,10,10,10,10,1
coast,shore,9.10,9,9.75,10,10,6,6,10,8,9.5,10,10,10,
football,soccer,9.03,9,9.9,9,10,8,7,9.5,10,8,10,9,9,
fuck,sex,9.44,10,9.75,10,10,8,9,9,10,9,10,8,10,10
journey,voyage,9.29,9,9.75,10,10,6,7,9.5,10,9.5,10,1
```

- split de http://alfonseca.org/eng/research/wordsim353.html

similarité	203 paires
relatedness	252 paires

À propos du split de *[Agirre et al., 2009]*

- 103 paires en commun dans les deux fichiers :

```
more wordsim_relatedness_goldstandard.txt \
     wordsim_similarity_goldstandard.txt | \
     sort | uniq -c | awk '$1 == 2 {print $0}' | \
     sort -k4,4gr | head -n 5+
```

```
1      2 cup          food    5,00
2      2 monk         oracle  5,00
3      2 journal      association     4,97
4      2 car          flight  4,94
5      2 street       children        4,94
```

- lire l'article pour plus d'information . . .

Rubenstein and Goodenough

http://www.cs.cmu.edu/~mfaruqui/word-sim/EN-RG-65.txt

```
1  gem jewel 3.94
2  midday noon 3.94
3  automobile car 3.92
4  cemetery graveyard 3.88
5  cushion pillow 3.84
6  ...
7  autograph shore 0.06
8  fruit furnace 0.05
9  noon string 0.04
10 rooster voyage 0.04
11 cord smile 0.02
```

rali

Word2Vec datasets (analogies)

http://word2vec.googlecode.com/svn/trunk

questions-words.txt 19544 équations (avec solution) :

```
: capital-common-countries
Berlin Germany Islamabad Pakistan
Baghdad Iraq Oslo Norway
: capital-world
Algiers Algeria Belmopan Belize
Abuja Nigeria Doha Qatar
: currency
Armenia dram India rupee
Angola kwanza Denmark krone
: city-in-state
Philadelphia Pennsylvania Atlanta Georgia
Chicago Illinois Stockton California
: family
father mother husband wife
brother sister prince princess
: gram1-adjective-to-adverb
cheerful cheerfully efficient efficiently
apparent apparently free freely
: gram2-opposite
clear unclear impressive unimpressive
aware unaware efficient inefficient
: gram3-comparative
bright brighter simple simpler
big bigger bright brighter
: gram4-superlative
bright brightest weird weirdest
big biggest dark darkest
: gram5-present-participle
debug debugging code coding
dance dancing enhance enhancing
: gram6-nationality-adjective
Australia Australian India Indian
Albania Albanian Sweden Swedish
: gram7-past-tense
describing described playing played
dancing danced vanishing vanished
: gram8-plural
bottle bottles man men
bird birds bottle bottles
: gram9-plural-verbs
enhance enhances play plays
describe describes increase increases
```

Word2Vec datasets (analogies)

http://word2vec.googlecode.com/svn/trunk

questions-phrases.txt 3218 équations (avec solution) :

```
: newspapers
    Boston Boston_Globe Salt_Lake Salt_Lake_Tribune
    Baltimore Baltimore_Sun Cincinnati Cincinnati_Enquirer
: ice_hockey
    Calgary Calgary_Flames Ottawa Ottawa_Senators
    Boston Boston_Bruins Chicago Chicago_Blackhawks
: basketball
    Dallas Dallas_Mavericks Boston Boston_Celtics
    Boston Boston_Celtics New_York New_York_Knicks
: airlines
    Finland Finnair Spain Spanair
    Austria Austrian_Airlines Turkey Turkish_Airlines
: people-companies
    Samuel_J._Palmisano IBM Jeff_Bezos Amazon
    Eric_Schmidt Google Larry_Ellison Oracle
```

rali

À propos des datasets Word2Vec

`http://word2vec.googlecode.com/svn/trunk`

- 5031 analogies impliquant la relation capitale_of, mais seulement 118 paires de pays / capitale . . .
- 870 analogies impliquant la relation plural-verbs, mais seulement 30 paires de verbes . . .
  - 27 avec la relation `+s` ; ex : eat / eats
  - et seulement 3 avec la relation `+es` : go / goes, search / searches et vanish / vanishes
  - **note** : [go: goes :: eat : eats] n'est pas une analogie formelle
- 1560 analogies impliquant la relation past-tense, mais seulement 40 paires de verbes . . .
  - ex : decreasing / decreased, knowing / knew
  - dont la moitié (23) sont des verbes irréguliers ; ex : sitted / sat et sleeped / slept

rali

MSR Sentence Completion dataset

http://research.microsoft.com/en-us/projects/scc

- 1040 questions comme celles-ci :

26 It may be that the solution of the one may _____ to be the solution of the other.

a) prove
b) learn
c) choose
d) forget
e) succumb

27 We shall just be in time to have a little _____ with him.

a) breakfast
b) elegance
c) garment
d) basket
e) dog

MSR Sentence Completion dataset

http://research.microsoft.com/en-us/projects/scc

- 1040 questions comme celles-ci :

26 It may be that the solution of the one may prove to be the solution of the other.

a) prove
b) learn
c) choose
d) forget
e) succumb

27 We shall just be in time to have a little breakfast with him.

a) breakfast
b) elegance
c) garment
d) basket
e) dog

MEN

http://clic.cimec.unitn.it/~elia.bruni/MEN.html

- 3000 paires de mots avec un score (entre 0 et 50) indiquant leur *relatedness* (200 TRAIN, 100 TEST)

cat-n feline-j	48	gun-n stair-n	2
copper-n metal-n	48	military-j tomato-n	2
flower-n garden-n	48	angel-n gasoline-n	1
grass-n lawn-n	48	feather-n truck-n	1
pregnancy-n pregnant-j	48	festival-n whisker-n	1

- paires vérifiant des contraintes de fréquences dans différentes ressources et échantillonnées de manière à générer une tâche balancée
- évaluation faite par comparaison binaire à 50 autres paires sur Mecanical Turk
- **tâche :** trouver les paires les plus similaires (ranking)

BLESS

https://sites.google.com/site/geometricalmodels/shared-evaluation

- 26 554 lignes comme celle-ci :

  ```
  frog-n  amphibian_reptile  attri  green-j
  ```

  qui indique qu'une grenouille est un amphibien ou un reptile et qu'un de ses attributs est d'être verte (gasp).

- 200 concepts (reptiles, mammifères, oiseaux, fruits, insects, armes, etc.) , 5 relations (attribut, co-hyponyme, hyperonyme, meronyme, événement)

  On average, there are 14 attri, 18 coord, 19 event, 7 hyper and 15 mero relata per concept.

- 14 400 concept-relation-relatum + 12 154 control tuples (random)
- une tâche suggérée

BLESS/frog

https://sites.google.com/site/geometricalmodels/shared-evaluation

mero	blood-n bone-n eye-n foot-n gill-n head-n leg-n lung-n poison-n skin-n tongue-n wart-n
event	breathe-v catch-v chase-v croak-v die-v drink-v eat-v hop-v kill-v lay-v leap-v live-v sit-v swim-v
attri	amphibious-j aquatic-j brown-j colorful-j edible-j green-j little-j loud-j noisy-j old-j poisonous-j small-j ugly-j wild-j
coord	alligator-n bullfrog-n lizard-n snake-n toad-n turtle-n
hyper	amphibian-n animal-n beast-n chordate-n creature-n vertebrate-n

+ 50 relations aléatoires

```
frog-n amphibian_reptile random-n doyen-n
```

SemEval 2010

Cause-Effect (1003)

```
He had chest pains and <e1>headaches</e1> from
<e2>mold</e2> in the bedrooms.
```

Component-Whole (941)

```
The <e1>provinces</e1> are divided into
<e2>counties</e2> (shahrestan), and ...
```

Entity-Destination (845)

```
The prosecutors carried <e1>guns</e1> into the
<e2>court room</e2>.
```

Product-Producer (717)

```
Next the <e1>soldiers</e1> put up a <e2>wall</e2>
of stakes on the pile of dirt.
```

+ Entity-Origin (716), Member-Collection (690), Message-Topic (634), Content-Container (540), Instrument-Agency (504)

- **Tâche** : identifier la relation de 2717 paires de noms marquées dans des phrases

SemEval 2012 (pilot) / 2013 (main task)

- sur une échelle de 0 à 5, noter la similarité de paires de phrases :

MT	When we are faced with a potential risk, it is important to apply the precautionary principle.
5	When we are faced with a potential risk, it is important to put into practice the precautionary principle.
MT	This rule can be applied equally to all recreational but dangerous drugs when no one but the person consuming the drug is likely to be affected.
4.8	This rule can, moreover, apply to all drugs récréationnelles but dangerous provided that the consumer of drugs is the only run the risks.
MSR-VID	A man stares out a window.
4.8	A man looks out the window.
MRS-PAR	But at age 15, she had reached 260 pounds and a difficult decision : It was time to try surgery.
4.2	But at the age of 15, she weighed a whopping 117kg and came to a difficult decision : it was time to try surgery.

SemEval 2014 / Task3

- Phrase2Word (TRAIN : 500, TEST : 500) — échelle $[0, 4]$

4	the structural and functional unit of all known living organisms	DNA	Lexicographic
4	traditionally the land on which a college or university and related institutional buildings are situate	campus	Lexicographic
4	watering hole	pub	Idiomatic
4	wear out one's welcome	exhaust	Idiomatic
4	whole ball of wax	total	Idiomatic
	⋮		
0	to make from scratch	breath	Idiomatic
0	what a vivid imagination	insipid	Newswire
0	workouts to strengthen abdominal muscles	fractions	Search

SemEval 2014 / Task3

- Sentence2Phrase (TRAIN : 500, TEST : 500) — échelle $[0, 4]$

4	We completely agree and there's no daylight between us on the issue.	complete agreement	Idiomatic
4	do you now where i can watch free older movies online without download ?	streaming vintage movies for free	CQA

- Paragraph2Sentence (TRAIN : 500, TEST : 500) — échelle $[0, 4]$

4	so i bought an item a couple weeks ago on eBay and i still havent even been notified that my item has been shipped yet.. what do i do ?	What should I do when my eBay purchase still hasn't shipped after two weeks ?	CQA

Bibliography I

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009).
A study on similarity and relatedness using distributional and wordnet-based approaches.
In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27.

Akbik, A., Michael, T., and Boden, C. (2014).
Exploratory relation extraction in large text corpora.
In *25th International Conference on Computational Linguistics*, pages 2087–2096.

Balasubramanian, N., Soderland, S., Mausam, and Etzioni, O. (2013).
Generating coherent event schemas at scale.
In *EMNLP'13*, pages 1721–1731.

rali

Bibliography II

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007).
Open information extraction from the web.
In *IN IJCAI*, pages 2670–2676.

Fader, A., Zettlemoyer, L., and Etzioni, O. (2014).
Open question answering over curated and extracted knowledge bases.
In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1156–1165.

Harris, Z. (1985).
Distributional structure.
The Philosophy of Linguistics, pages 26–47.

rali

Bibliography III

Mausam, Schmitz, M., Bart, R., Soderland, S., and Etzioni, O. (2012).
Open language learning for information extraction.
In *Joint EMNLP and CoNLL*, pages 523–534.

Mesquita, F., Schmidek, J., and Barbosa, D. (2013).
Effectiveness and efficiency of open relation extraction.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 447–457.

Min, B., Shi, S., Grishman, R., and Lin, C.-Y. (2012).
Ensemble semantics for large-scale unsupervised relation extraction.
In *Proceedings of the 2012 Joint EMNLP and CoNLL*, pages 1027–1037.

rali

Bibliography IV

Nakashole, N., Weikum, G., and Suchanek, F. (2012).
Patty : A taxonomy of relational patterns with semantic types.
In *Joint EMNLP and CoNLL*, pages 1135–1145.

Poon, H., Quirk, C., DeZiel, C., and Heckerman, D. (2014).
Literome : Pubmed-scale genomic knowledge base in the cloud.
Bioinformatics, 30(19) :2840–2842.

rali