Apprentissage de représentations jointes de mots et d'entités : expériences sur une collection textuelle liée à une base de connaissance

Jose G. Moreno

IRIT

12/11/2018

LIMSI

- Romain Beaumont*
- Eva D'hondt*
- Anne-Laure Ligozat
- Sophie Rosset
- Brigitte Grau
- Sanjay Kamath
- Yue Ma
- Rashedur Rahman
- Xavier Tannier*
- Cong Wang*

CEA

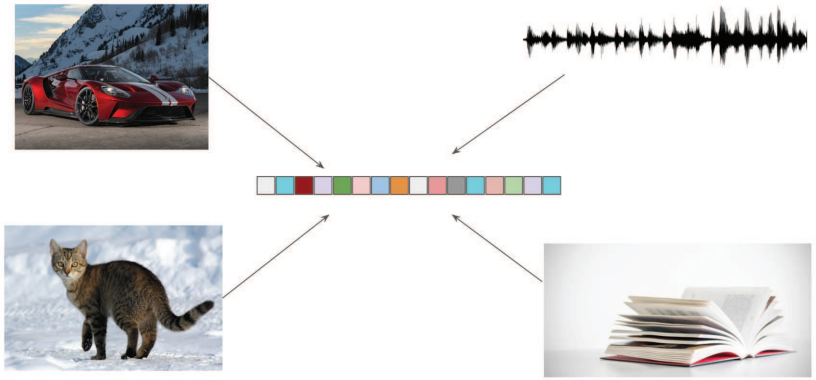
- Romaric Besançon

Représentation

Définition

- Générale : action de replacer devant les yeux de quelqu'un
- Notre travail : action de replacer devant les "yeux" d'une machine
- Une large variété de représentations possibles (histogrammes, graphes, etc), cependant la représentation vectorielle des objets a été récemment privilégié.

Représentation vectorielle (embeddings)

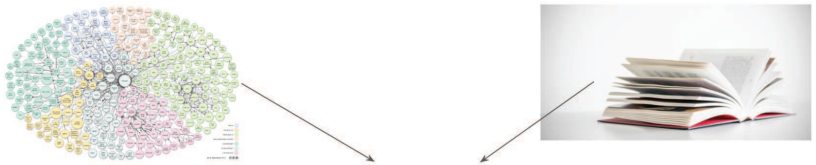


Avantages

- Comparaison facile entre éléments
- Représentation standard
- L'entrée pour des modèles plus élaborées

Représentation jointe

- Une représentation unique pour des objets de nature différents
- Les caractéristiques de similarité sémantique sont cohérents (par rapport à la source)



$$x \in \mathbb{R}^n$$

Représenter des unités du texte (mots) et des unités d'une base de connaissance (entités) dans une espace vectorielle unifié :

$$x_w^i, x_e^j \in \mathbb{R}^n$$

Défis pour construire une représentation jointe

Bases de connaissances

- Composé des triplets (sujet, prédicat, objet)
- Le sujet et l'objet peuvent être identifiés par des URI (entités)
- Les triplets sont des faits exprimés comme une information binaire et étiqueté entre les nœuds/entités du graphe

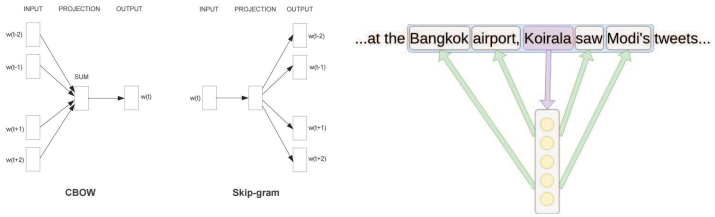
Documents texte

- Séquences des mots avec une cohérence sémantique et lexicale
- Des entités peuvent être mentionnées dans le texte mais avec différents mentions (Barack Obama, Obama, President Obama, etc.)
- Les mentionnes d'entités ne sont normalement pas répertoriés dans les documents textes

Apprentissage de représentation de mots

Plongements lexicaux

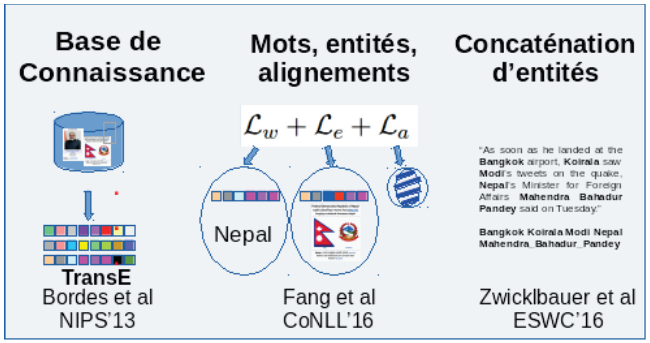
Fondés sur l'hypothèse distributionnelle



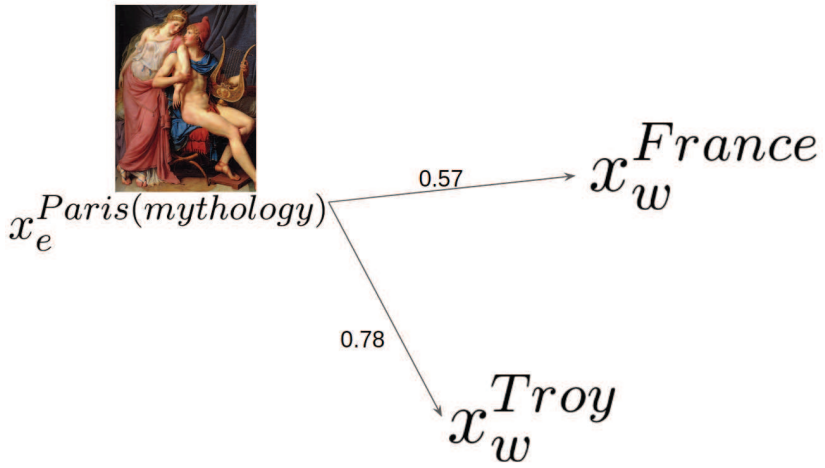
Apprentissage de représentation d'entités

Représentation d'entités

fondée également sur l'hypothèse distributionnelle



Avantages d'une représentation jointe



Ressourcés avec des informations jointes

Wikipédia

C'est une large collection textuelle qui contient des mentionnes d'entités avec une variabilité très riche pour chaque entité

Statistiques de Wikipédia

- (EN) Pages avec du contenu 5,749,083
- (EN) Toutes les pages (listes, redirections, etc.) 46,314,174
- 12 langues sur Wikipédia avec plus de 1M d'articles : Allemand, Espagnol, Français, Italien, Néerlandais, Japonais, Polonais, Portugais, Russe, Suédois, Vietnamien et Mandarin

Ressourcés avec des informations jointes

Page Wikipédia de Toulouse vue par un utilisateur

Toulouse is the capital of the French **department** of **Haute-Garonne** and of the **region** of **Occitanie**. The city is on the banks of the **River Garonne**, 150km from the **Mediterranean Sea**, 230km from the **Atlantic Ocean** and 680km from **Paris**.

Ressourcés avec des informations jointes

Page Wikipédia de Toulouse vue par un contributeur

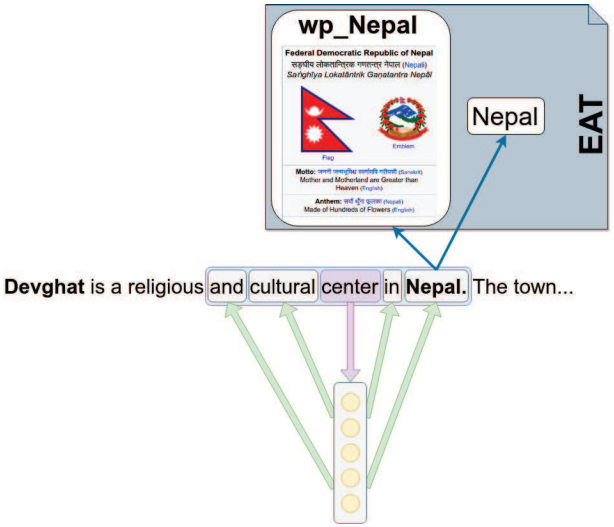
Toulouse is the capital of the French **[[departments of France|department]]** of **[[Haute-Garonne]]** and of the **[[regions of France|region]]** of **[[Occitanie]]**. The city is on the banks of the **[[Garonne|River Garonne]]**, 150km from the **[[Mediterranean Sea]]**, 230km from the **[[Atlantic Ocean]]** and 680km from **[[Paris]]**.

Modèle Extended Anchor Text (EAT)

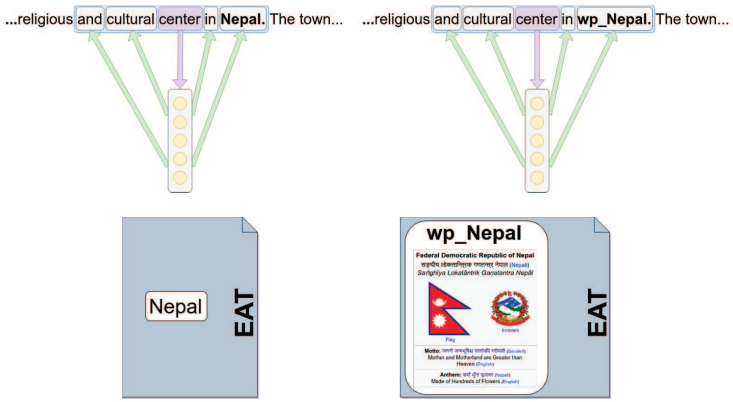
Caractéristiques

- EAT utilise un corpus annoté comme Wikipédia
- EAT permet une représentation jointe des mots et des entités avec des techniques classiques de plongement sémantique

Représentations mots vs entités

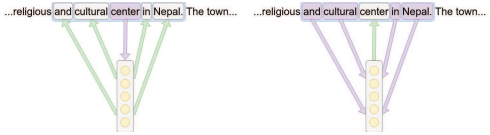


Représentations mots vs entités

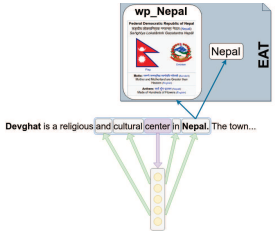


Caractéristiques du modèle EAT

- possibilité d'utiliser les configurations Skip-gram ou CBOW



- le texte d'ancrage permet de représenter conjointement mots et entités



- les vecteurs d'entités sont appris en utilisant leur contexte, et non indirectement (pas de concaténation ou alignement)

Caractéristiques techniques du modèle EAT

Implémentation

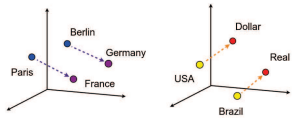
- Premiers versions en utilisant Gensim et Hyperwords
- Pré-traitement : EAT_wikiextractor.py + redirections
- Version stable sur TensorFlow
  - C++ (implémentation optimisé de skipgram) + Tensorflow https ://github.com/jgmorenof/EAT-tensorflow
  - 2 jours pour 5 époques sur 1 machine (30GB, 10 Cores, etc.)
  - Vecteurs disponibles sur demande (17 GB dézippé)

Caractéristiques techniques du modèle EAT

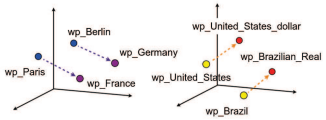
Les vecteurs des mots et d'entités

- 200 dimensions
- Vocab. de 5.2M = 1.8M d'entités et 3.4M de mots (tokens)

Ent Acc.	EAT-words	EAT	Google	Glove
$ALL_{Sem} - Family$	0.5870	0.6688	0.6503	0.6649



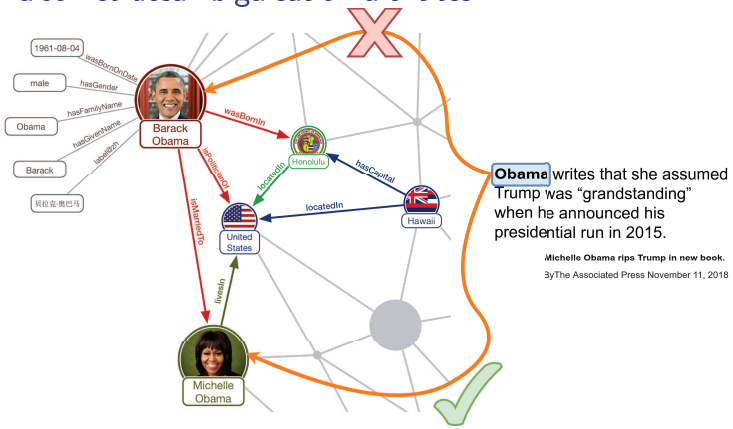
Word



Entity

Applications

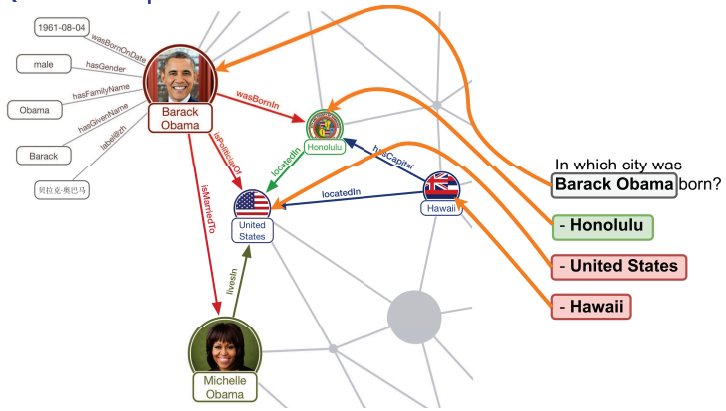
Liaison et désambiguïsation d'entités



Applications

Liaison et désambiguïsation d'entités

Question-réponse

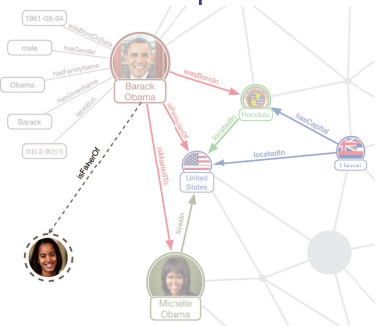


Applications

Liaison et désambiguïsation d'entités

Question-réponse

Validation de réponse



Since arriving on the island, **Barack**, Michelle, **Malia** and Sasha have gone on a hike at the Makiki Loop Hawaii Nature Center, seen President Obama's childhood school and dined at Morimoto restaurant.

The "Becoming" author and former President **Barack Obama** were able to welcome caughters, **Malia**, 20, and Sasha, 17, through IVF.

Applications

- **Liaison et désambiguïsation d'entités**
- Question-réponse
- Validation de réponse

Liaison et désambiguïsation d'entités

Définition

Relier des mentions d'entités d'un texte à un identifiant unique

- entité d'une base de connaissance (BC) si elle existe
- un regroupement de mentions sans référent (NIL)

Contexte

Données

Pour réaliser la tâche de liaison d'entité, on part de :

- base de connaissance
- des documents annotés en mentions à désambiguïser (+ gold)
  - un document est une page Web
  - une requête est une mention + 1 document

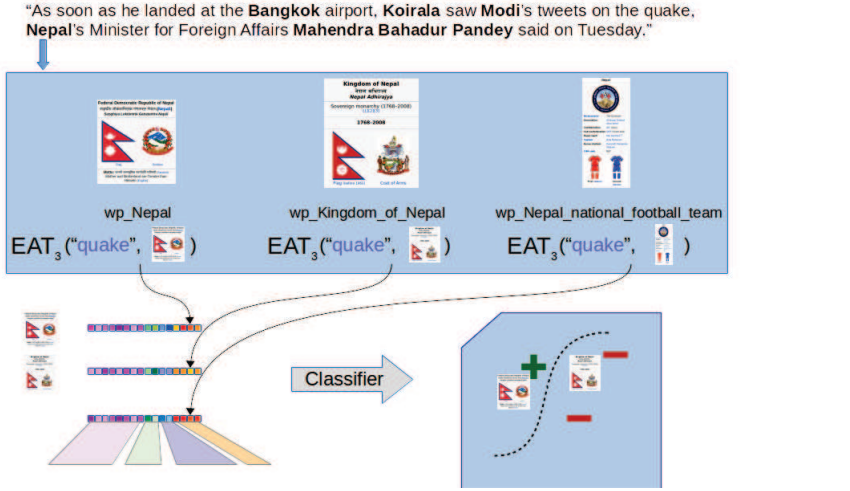
Quelles informations ?

- informations en provenance de la requête (mention + contexte)
- informations en provenance de la cible (BC)

Ces sources sont en général utilisées de façon indépendante
$\rightarrow$ hypothèse : leur combinaison devrait améliorer la liaison

Architecture

- Expansion de requête et génération de candidats
- Ordonnancement des candidats : traits initiaux + traits EAT



Traits initiaux (*baseline*)

Traits issus de l'étape de génération des candidats

- égalité entre les formes lexicales
  - mention d'entité et label d'entité dans la BC
  - mention d'entité et variante (alias ou traduction) d'un label
- inclusion de la mention de l'entité dans le label ou variantes
- distance de Levenshtein ≤ 2 entre mention et label ou variante

Traits contextuels

- similarité cosinus entre la page Wikipedia et le document contenant la mention
- similarité cosinus entre la page Wikipedia des entités en relation avec l'entité considérée et le document

Traits de popularité

- nombre de liens entrant sur la page Wikipedia
- nombre de visites de la page Wikipedia (2014)

Traits EAT

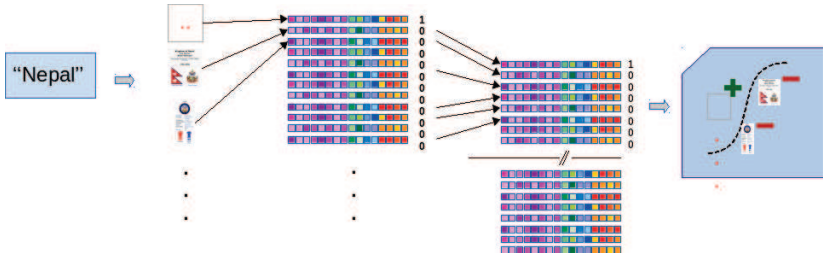
- EAT_1 : moyenne des similarités entité/paragraphe
- EAT_2 : similarité cosinus entre le vecteur moyen des mots du paragraphe et le vecteur de l'entité candidate
- EAT_3 : moyenne des k-meilleures similarités
- EAT_4 : similarité entre l'entité et la mention

Entraînement

Pour chaque mention

- génération des candidats
- les candidats corrects constituent les exemples positifs
- 10 candidats tirés au hasard constituent les exemples négatifs

Un modèle est appris par apprentissage supervisé

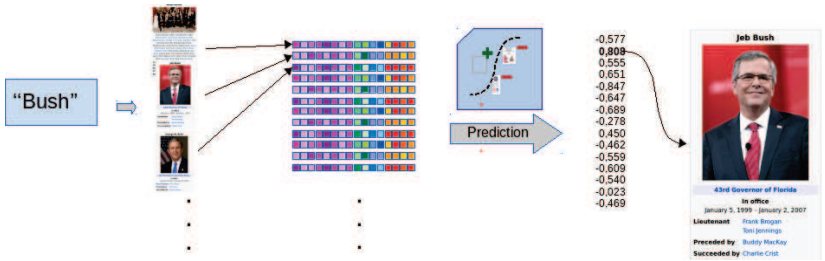


Sélection du meilleur candidat

Pour chaque requête

- génération des candidats
- chaque candidat est évalué avec le modèle
- le candidat avec le meilleur score positif est retourné

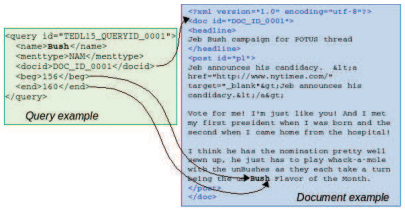
Si les prédictions sont négatives pour tous les candidats, alors la mention est considérée comme NIL



Expériences

Données : Jeu de test TAC EDL 2015

- Train
  - 168 documents
  - 12 175 requêtes (mentions)
  - 73,6% dans la BC
- Test
  - 167 documents
  - 13 175 requêtes (mentions)
  - 74,4% dans la BC



Expériences

Métriques

Précision et Rappel pour les *liens* et les *NIL*
avec t = système et r = référence

$$P(NIL) = \frac{N(e_t = NIL \wedge e_r = NIL)}{N(e_t = NIL)} \quad (1)$$

$$R(NIL) = \frac{N(e_t = NIL \wedge e_r = NIL)}{N(e_r = NIL)} \quad (2)$$

$$P(liens) = \frac{N(e_t = e_r \wedge e_t \neq NIL)}{N(e_t \neq NIL)} \quad (3)$$

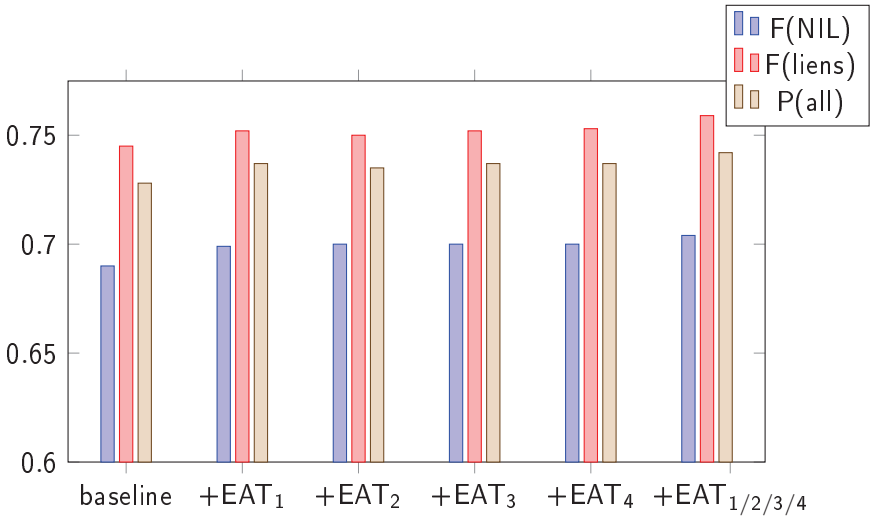
$$R(liens) = \frac{N(e_t = e_r \wedge e_t \neq NIL)}{N(e_r \neq NIL)} \quad (4)$$

$$P(all) = \frac{N(e_t = e_r)}{N(e_t)} \quad (5)$$

Résultats de la génération des candidats

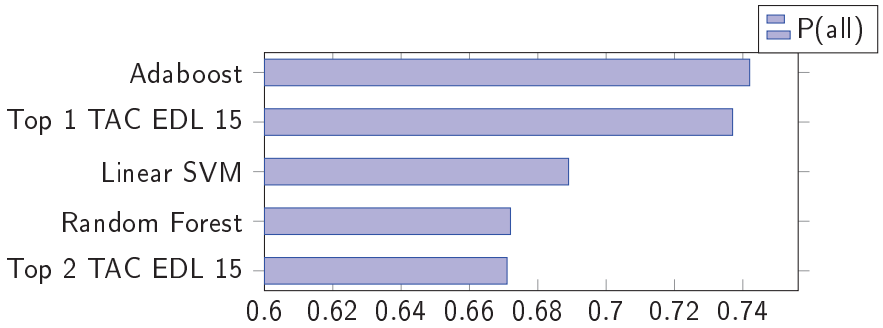
Tous	$\lvert C\rvert$	$\lvert C_{NIL}\rvert$	C_{AVG}	Rappel(C)	P(all)
train	6 843 513	781	562,1	95,60 %	
test	8 339 648	499	613,8	94,19 %	0,646
Typage	$\lvert C\rvert$	$\lvert C_{NIL}\rvert$	C_{AVG}	Rappel(C)	P(all)
train	3 179 795	952	261,2	92,43 %	
test	3 810 382	626	280,4	90,36 %	0,680
Typage-sim	$\lvert C\rvert$	$\lvert C_{NIL}\rvert$	C_{AVG}	Rappel(C)	P(all)
train	1 723 470	952	141,6	90,27 %	
test	1 921 577	625	141,4	87,95 %	0,714

Résultats : traits



- chaque trait améliore le système initial
- la combinaison de tous les traits donne les meilleurs résultats et est supérieure à l'état de l'art

Résultats : classifieurs



- trois différents algorithmes de classification : Adaboost, SVM linéaire, Random Forest
- meilleurs résultats obtenus avec Adaboost

Applications

- Liaison et désambiguïsation d'entités
- **Question-réponse**
- Validation de réponse

Question-réponse

Question originale

- Q : What is the main executive body of the EU ?
- P : The European Commission is the main executive body of the European Union.
- A : European Commission

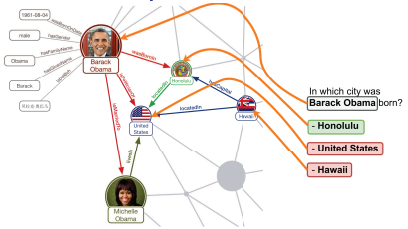
Question avec des entités

- Ent.Q : What is the main executive body of the **wp_European_Union** ?
- Ent.P : The **wp_European_Commission** is the main executive body of the **wp_European_Union**.
- Ent.A : **wp_European_Commission**

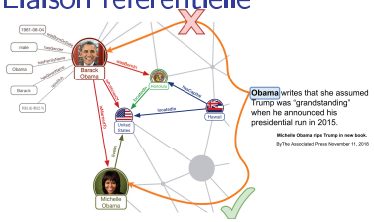
Résumé

Modèle EAT

Question-réponse



Liaison référentielle



Validation de relation