Traiter des données relationnelles grâce à l'apprentissage automatique

Antoine Bordes

CNRS - Heudiasyc - UTC

Travaux financés le projet ANR EVEREST.

PFIA – CAp – Lille – 5 juillet 2013

Avec Université de Montréal, IDIAP, INRIA, Google et Xerox.

Préambule

"On va vivre dans des bases de connaissance."

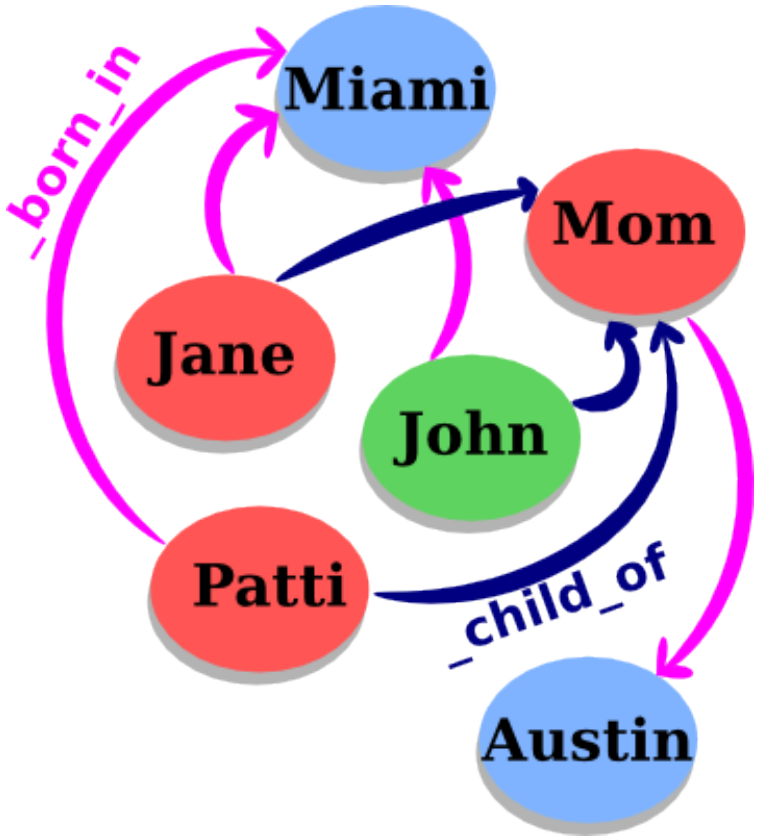
"Dans un monde ouvert, imprécis et incohérent."

Serge Abiteboul
PFIA, le 4/7/13.

Menu

Les données relationnelles

- Les données sont structurées selon un graphe.
- Chaque noeud = une entité.
- Chaque arc = une relation.
- Une relation = (*sub*, *rel*, *obj*) :
  - *sub* =*sujet*,
  - *rel* = *type de relation*,
  - *obj* = *objet*.
- Noeuds sans caractéristiques.



Exemples de données relationnelles

Ce type de données structurées est très courant :

- Réseaux sociaux : Facebook, LinkedIn, etc.
- En bio-informatique : réseaux de régulation, etc.
- En marketing/recommandation : traces, notes, etc.
- Bases de connaissances : Freebase, Yago, WordNet, etc.

- De façon général, toute donnée au format RDF.
- Format simple pour stocker des données non-structurées.

Caractéristiques de telles données

Les caractéristiques varient selon les tâches.

Néanmoins, en général, ces données sont :

- de grandes dimensions ($10^5 - 10^8$ noeuds, $10^7 - 10^9$ relations),
- creuses (peu de liens valides),
- incomplètes (relations et noeuds manquants),
- bruitées/incohérentes (relations et noeuds incorrects),
- hétérogènes (propriétés varient bcp selon le type de relation).

Problématiques

On cherche donc à traiter les données relationnelles pour :

- les compléter (ajout de relations et/ou noeuds),
- les débruiter/réparer (suppression de relations et/ou noeuds),
- les résumer (interprétation, visualisation),
- les fusionner entre elles (projet LinkedData),
- les rendre plus facilement manipulables et utilisables dans d'autres applications (en TALN, en recherche d'info., en vision).

⟶ **valoriser les données relationnelles.**

Menu

Les bases de connaissances

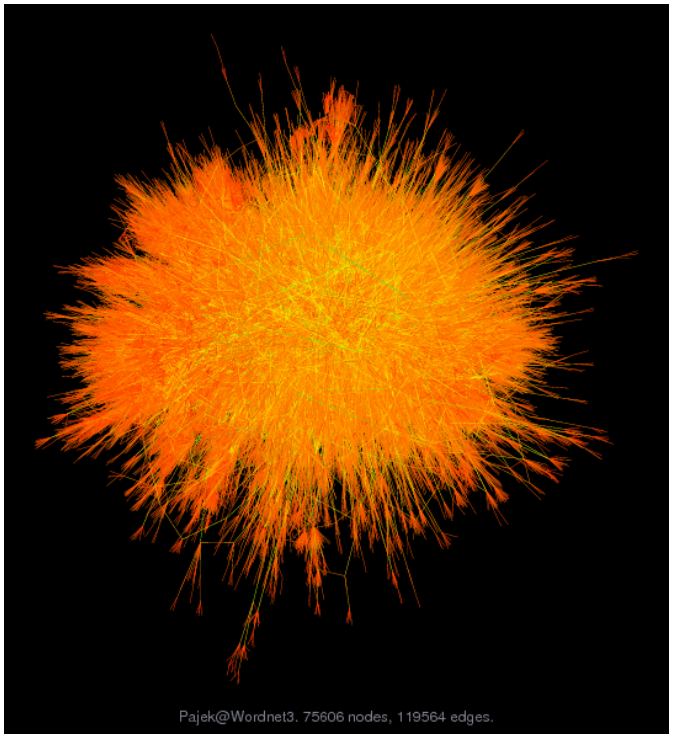
Nos travaux sont axés sur les bases de connaissances (BCs) et en particulier, sur les bases de (très) grandes tailles.

- **But :** améliorer des tâches en apportant des connaissances.
  - recherche d'information (Google Knowledge Graph, Kindle),
  - question-réponse (IBM Watson),
  - recommandation (Netflix+IMDB),
  - biologie (veille scientifique)
  - TALN, etc.
- **Exemples :** (libres d'accès)
  - Génériques : Freebase, YAGO, DBpedia, (Open)Cyc, etc.
  - Bio-informatique : GeneOntology, IntAct, UniprotKB, etc.
  - Linguistique : WordNet.
  - Géographie : GeoNames.
  - Cinéma : IMDB.

Exemple 1 : WordNet

- WordNet : graphe-dictionnaire où chaque noeud est un sens.

- Très utilisé en TALN.
- Caractéristiques :
  - 117k d'entités ;
  - 20 types de relations ;
  - 500k relations.
- Exemples :
  (car_NN_1, _has_part, _wheel_NN_1)
  (score_NN_1, _is_a, _rating_NN_1)
  (score_NN_2, _is_a, _sheet_music_NN_1)



Exemple 2 : Freebase

- Freebase : gigantesque BC collaborative (donc bruitée).

- Une partie du Knowledge Graph de Google.
- Caractéristiques :
  - 80M d'entités ;
  - 23k types de relations ;
  - 1.2B relations.
- Exemples :
  (Serge Abiteboul, _academic_advisor, Seymour Ginsburg)
  (Lille, _contained_by, Nord)
  (Machine Learning, _subdiscipline, Artificial Intelligence)

Objectifs

Notre but : faciliter la manipulation de ces grandes BCs.

- Ajouter de la connaissance :
  - Visualisation ;
  - Prédiction de liens ;
  - Résumé (dans le futur).

- Connecter le texte et les BCs :
  - Compléter Freebase grâce au texte ;
  - Désambiguïser le texte avec à WordNet.

Notre moyen : apprentissage statistique de représentations.

Menu

Apprentissage statistique relationnel

- **Cadre :**
  - n_s sujets $\{sub_i\}_{i \in [\![1;n_s]\!]}$
  - n_r types de relation $\{rel_k\}_{k \in [\![1;n_r]\!]}$
  - n_o objets $\{obj_j\}_{j \in [\![1;n_o]\!]}$

  → Pour nous, $n_s = n_o = n_e$ et $\forall i \in [\![1;n_e]\!], sub_i = obj_i$.

  - Une relation existe pour (sub_i, rel_k, obj_j) si $rel_k(sub_i, obj_j) = 1$
- **But :** On veut modéliser, à partir des données,

$$\mathbb{P}[rel_k(sub_i, obj_j) = 1]$$

(équivalent à approximer le tenseur binaire $\mathbf{X} \in \{0,1\}^{n_s \times n_o \times n_r}$)

Problème d'apprentissage

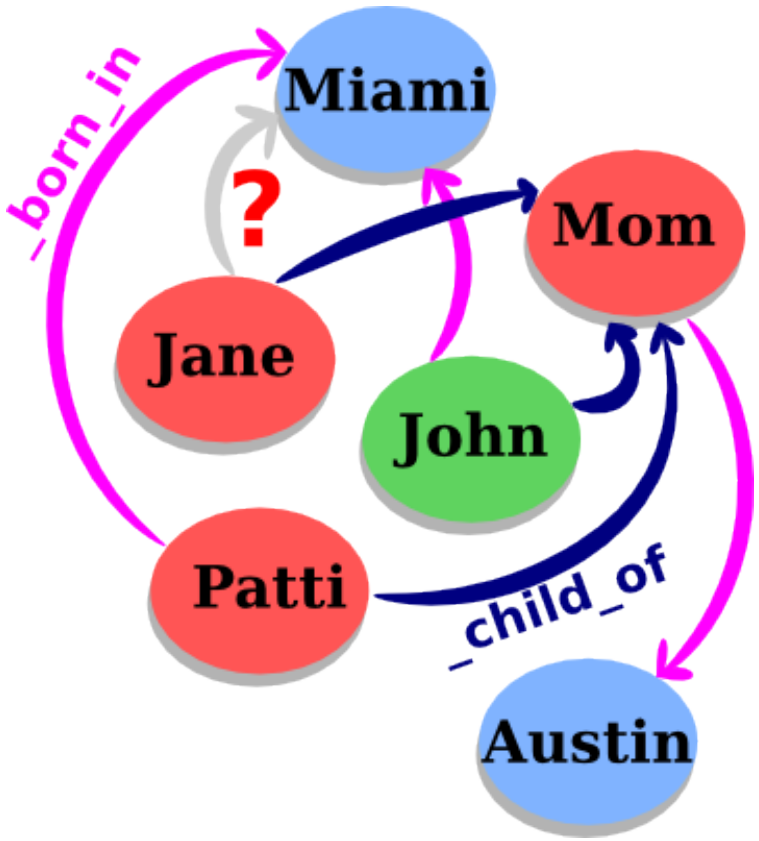
- **Principes** :
  - chercher les régularités ;
  - condenser les données ;
  - définir des mesures de similarité.

- **Caractéristiques** :
  - forte structure sur les exemples d'apprentissage ;
  - grande dimensions ;
  - données creuses : classes déséquilibrées.
  - données manquantes.

Généralisation

Compléter les BC à partir des informations déjà présentes.

A partir des probabilités connues,
→ proba. d'une relation inconnue.

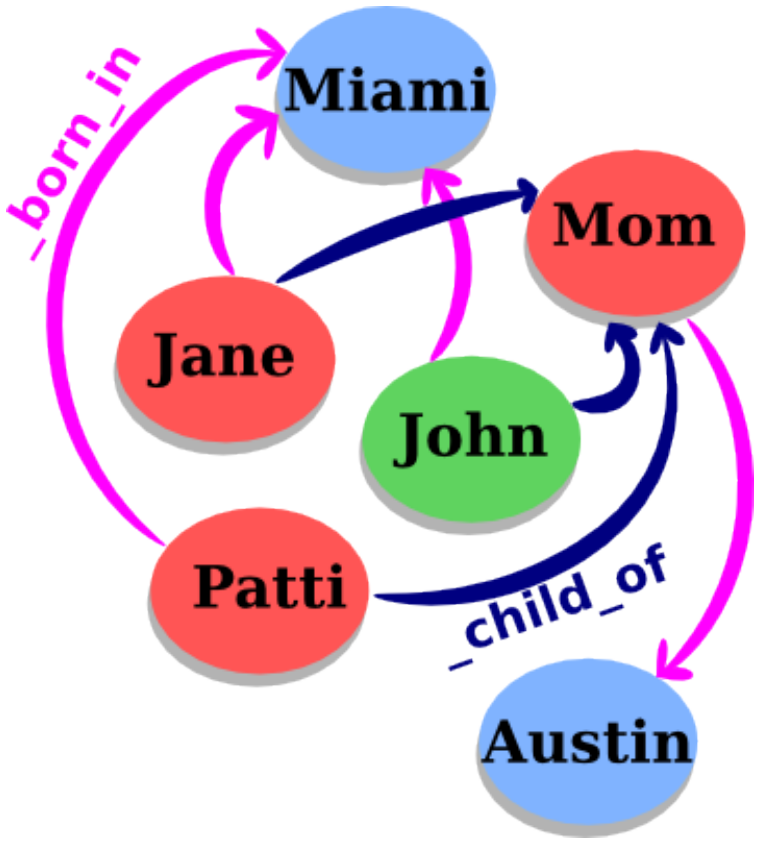


Généralisation

Compléter les BC à partir des informations déjà présentes.

A partir des probabilités connues,
→ proba. d'une relation inconnue.

→ *classification collective*.
→ basée sur des similarités.

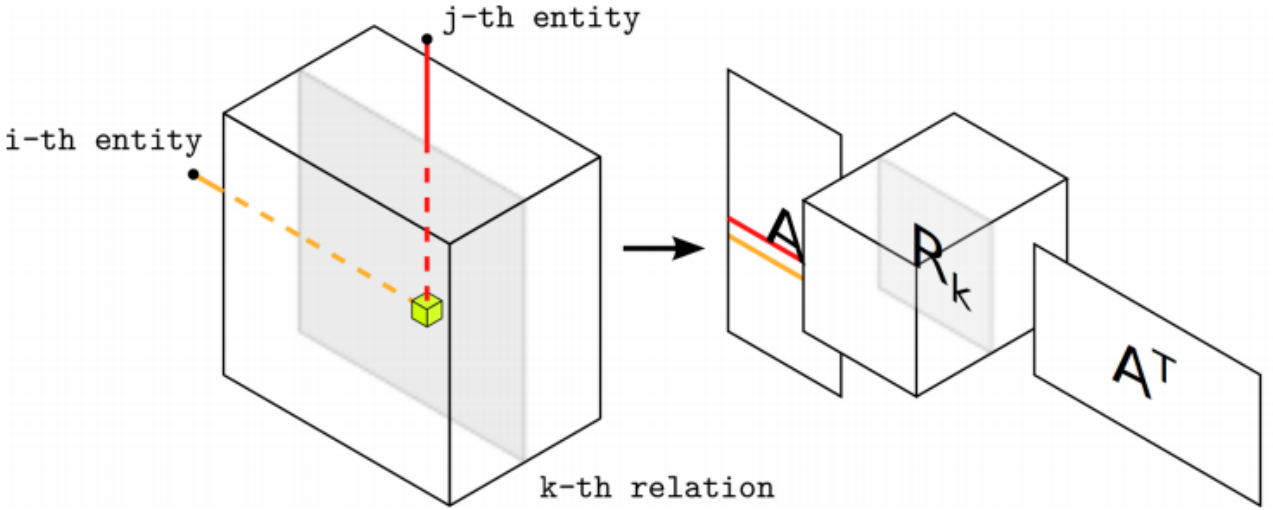


Méthodes utilisées

- Factorisation de tenseurs (e.g. (Harshman et al., 94)).
- Avec modélisation des valeurs manquantes (e.g. (Gao et al., 11))
- *Markov-logic Networks* (e.g. (Kok et al., 07))
- *Clustering* d'entités/relations (e.g. BCTF (Sutskever et al., 10)).
- *Spectral clustering* pour les multi-graphs (Dong et al., 11).
- Factorisation collective de matrices (e.g. (Nickel et al., 11)).

Factorisation de collective de matrices (RESCAL)

- $\forall k \in [\![1; n_r]\!], \mathbf{R}_k \in \mathbb{R}^{d \times d}$ et $\mathbf{A} \in \mathbb{R}^{n_e \times d}$.
  (proche de DEDICOM (Harshman, 78)).



- App. de **A** et **R** par reconstruction (moindres carrés alternés) :

$$\min_{\mathbf{A},\mathbf{R}} \frac{1}{2}\left(\sum_k ||\mathbf{X}_k - \mathbf{A}\mathbf{R}_k\mathbf{A}^\top||_F^2\right) + \lambda_A||\mathbf{A}||_F^2 + \lambda_R \sum_k ||\mathbf{R}_k||_F^2$$

Factorisation de collective de matrices (RESCAL)

- **Avantages** :
  - bonne qualité de prédiction de lien, en pratique.
  - rapide tant que les données tiennent en mémoire.
  - bonne utilisation de la parcimonie des données.

- **Inconvénients** :
  - "beaucoup" de paramètres ($d >> 1$).
  - apprentissage coûteux sur de très grandes tailles.
  - peu flexible (ajout de nouveaux éléments).
  - critère d'apprentissage peu adapté.
  - pas de gestion des données manquantes.

Menu

Notre approche

Deux idées principales :

1. Modèles basés sur des représentations vectorielles de faible dimension pour les entités et les relations apprises suivant un critère de similarité.

2. Apprentissage stochastique avec un sous-échantillonnage des relations non-observées.

Apprentissage de représentations

- Sujets et objets sont représentés par des vecteurs de $\mathbb{R}^d$.
  - $\{sub_i\}_{i\in[\![1;n_s]\!]} \quad \rightarrow \quad [\mathbf{s}^1,\ldots,\mathbf{s}^{n_s}] \in \mathbb{R}^{d\times n_s}$
  - $\{obj_j\}_{j\in[\![1;n_o]\!]} \quad \rightarrow \quad [\mathbf{o}^1,\ldots,\mathbf{o}^{n_o}] \in \mathbb{R}^{d\times n_o}$

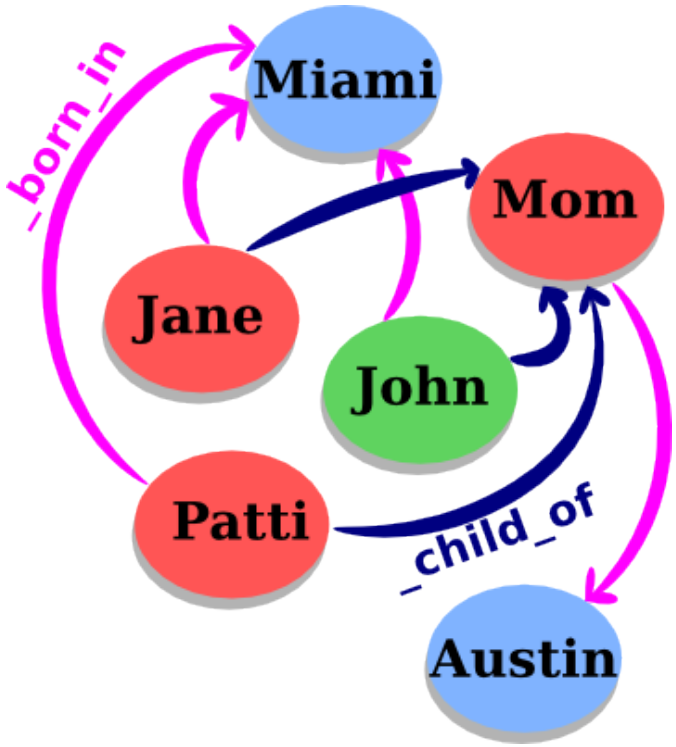
  Pour nous, $n_s = n_o = n_e$ et $\forall i \in [\![1;n_e]\!], \mathbf{s}_i = \mathbf{o}_i$.

- Types de rel. = opérateurs de similarités entre sujets/objets.
  - $\{rel_k\}_{k\in[\![1;n_r]\!]} \quad \rightarrow \quad$ operateurs $\{\boldsymbol{r}_k\}_{k\in[\![1;n_r]\!]}$

- App. de similarités dépendantes de $rel \rightarrow d(sub, rel, obj)$.
  (on revient à des proba. avec une fct de transfert.)

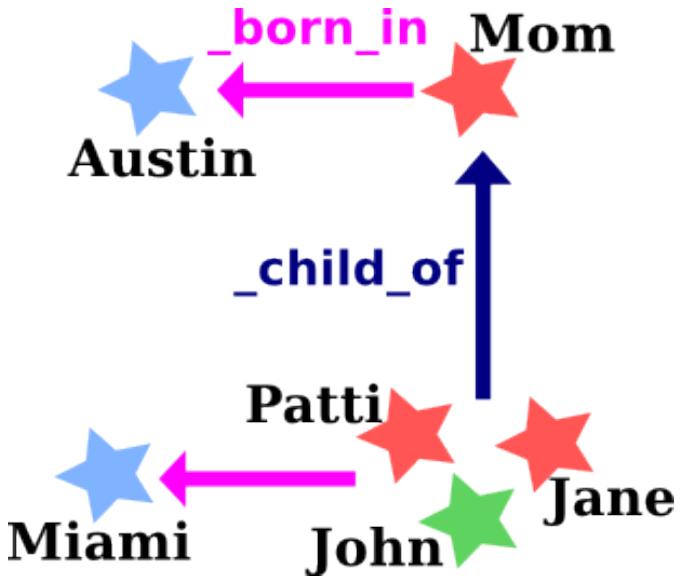
Modéliser les relations comme des translations

Intuition : on voudrait que $\boldsymbol{s} + \boldsymbol{r} \approx \boldsymbol{o}$.



Modéliser les relations comme des translations

Intuition : on voudrait que $\boldsymbol{s} + \boldsymbol{r} \approx \boldsymbol{o}$.



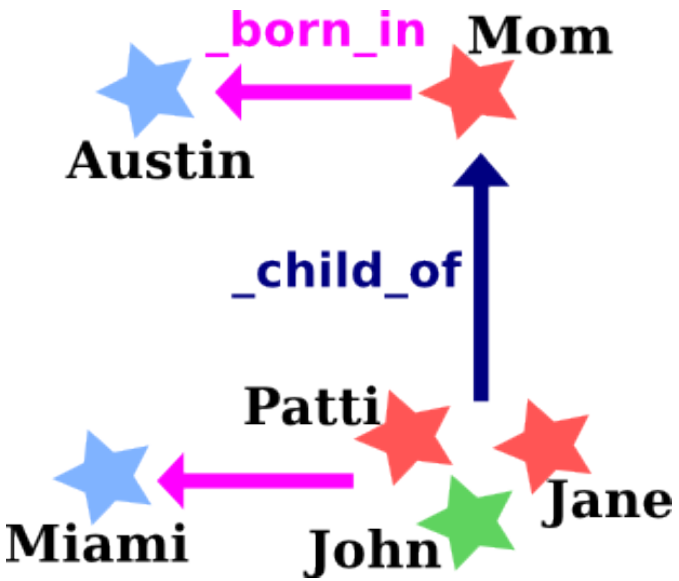
Modéliser les relations comme des translations

Intuition : on voudrait que $\boldsymbol{s} + \boldsymbol{r} \approx \boldsymbol{o}$.

On défini la mesure de similarité :

$$d(sub, rel, obj) = ||\boldsymbol{s} + \boldsymbol{r} - \boldsymbol{o}||_2^2$$

On apprend $\boldsymbol{s}$,$\boldsymbol{r}$ et $\boldsymbol{o}$ qui la vérifient.



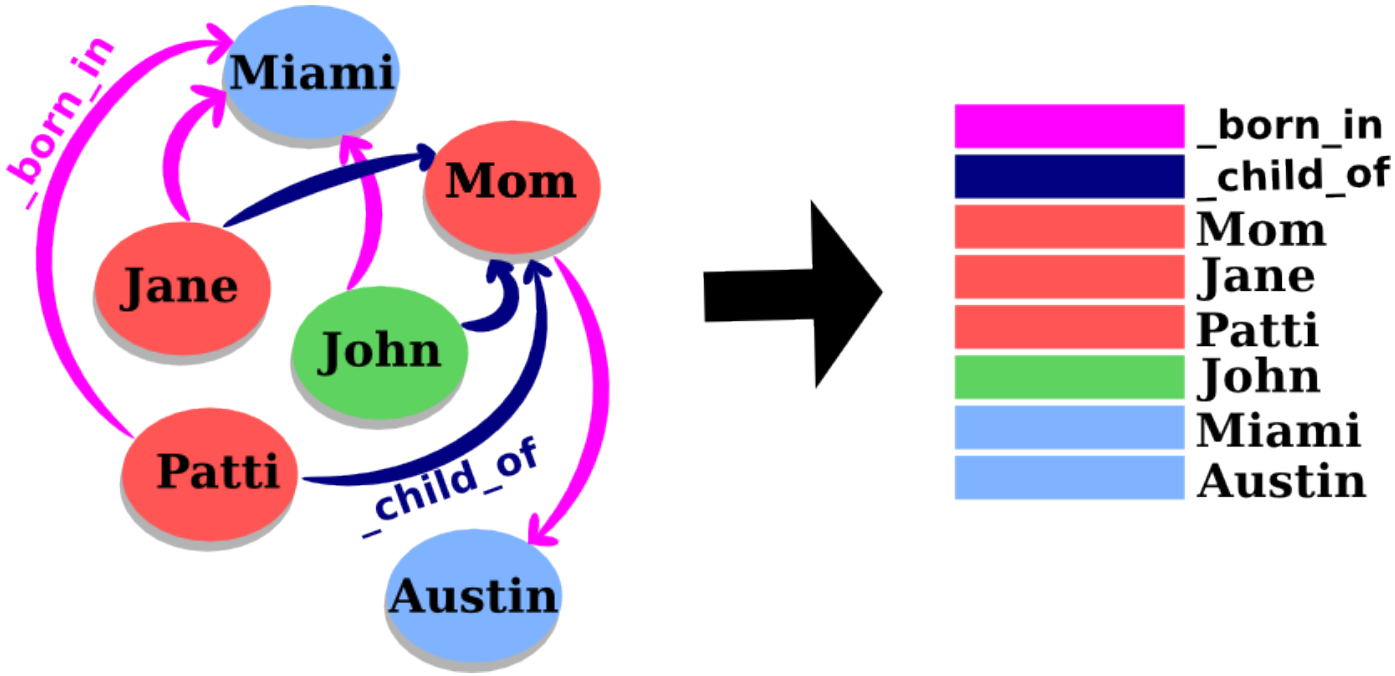
Apprentissage stochastique

- Apprentissage par descente de gradient stochastique : une relation observée (vraie ?) après l'autre.

- Pour chaque relation de l'ensemble d'apprentissage :
  1. on sous-échantillonne des rel. non-observées (fausses ?)
  2. on vérifie si la similarité de la vraie relation est inférieure.
  3. si non, on met à jour les paramètres des relations concernées.

- Critère d'arrêt : performance sur un ensemble de validation.

En résumé



Propriétés :

- peu coûteux en paramètres ;
- représentations apprises selon une mesure de similarité ;
- apprentissage simple à mettre oeuvre ;
- ajout facilité de nouveaux éléments.

Menu

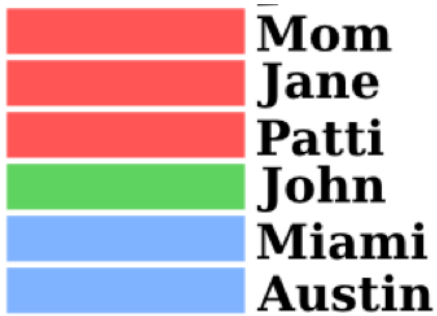
Sous-ensembles de Freebase

- **Données pour les expériences** :

	Entités (n_e)	Rel. (n_r)	Ex. App	Ex. Valid.	Ex. Test
Freebase15k	14,951	1,345	483,142	50,000	59,071
Freebase1M	1×10^9	23,382	17.5×10^9	50,000	177,404

- **Conditions expérimentales** :
  - Dimension des représentations : 50.
  - Durées d'apprentissage :
    - sur Freebase15k : ≈5h (sur 1 seule machine),
    - sur Freebase1M : ≈1j (Map/Reduce sur 16 machines).

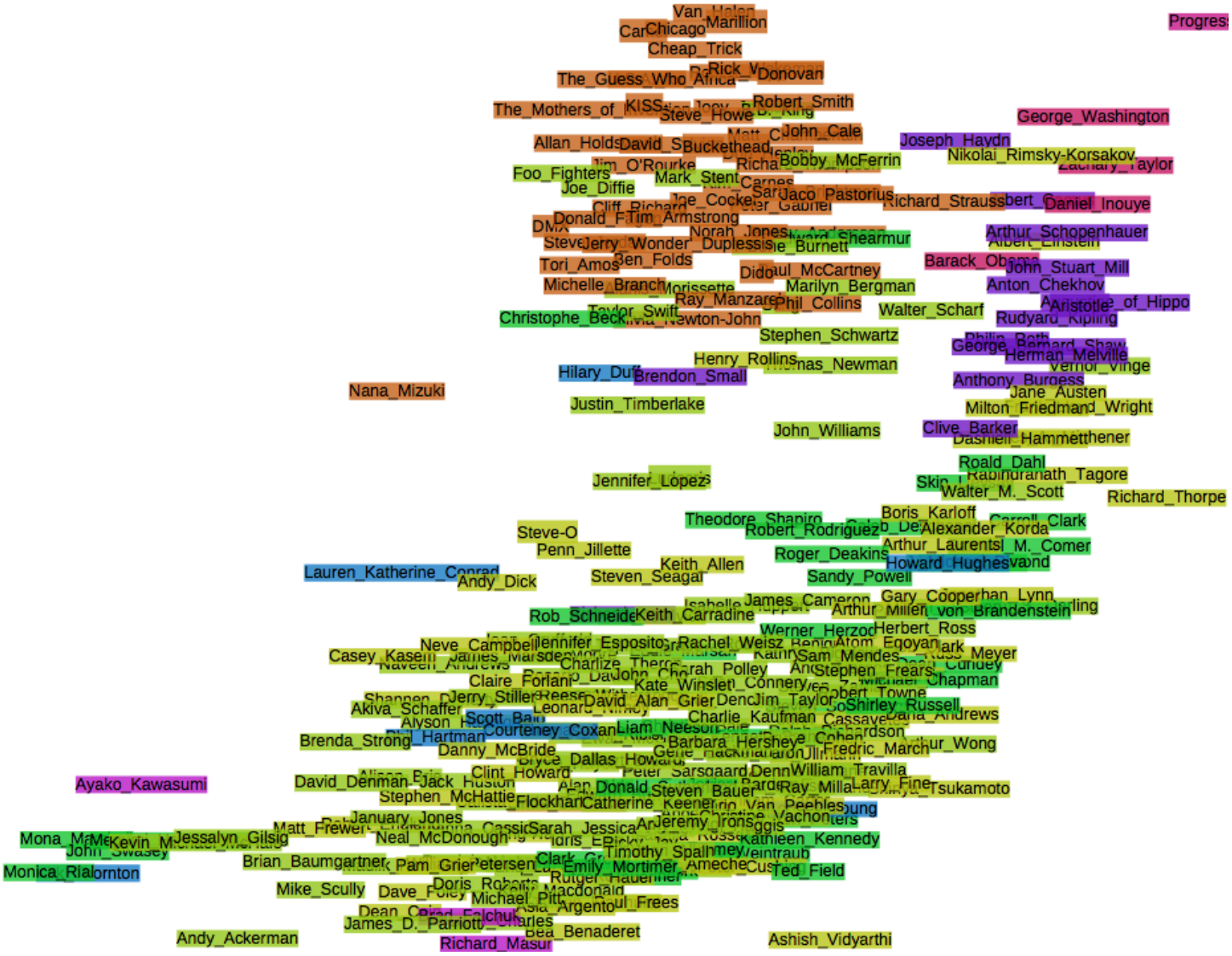
Visualisation de 1000 entités



Projection en 2D des similarités entre vecteurs représentant 1,000 entités de Freebase15k en utilisant Gephi.

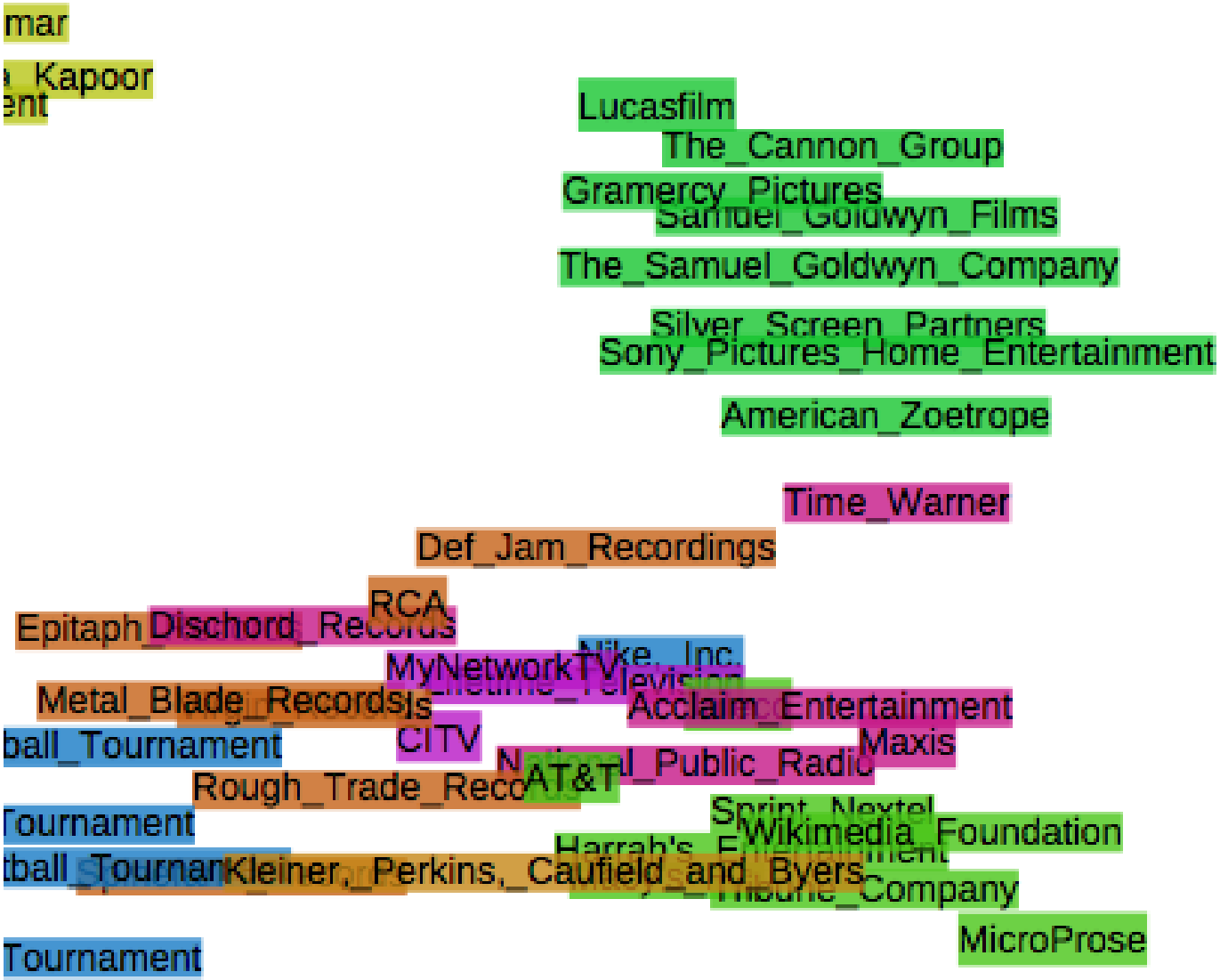
Visualisation de 1000 entités

Visualisation de 1000 entités - Zoom 1



Visualisation de 1000 entités - Zoom 2

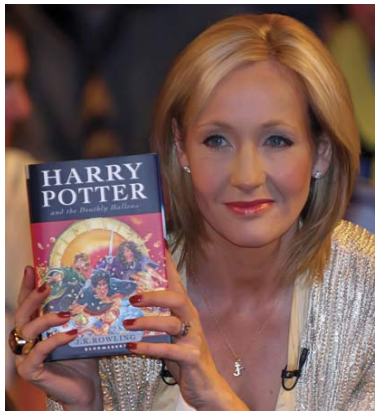
Visualisation de 1000 entités - Zoom 3



Prédiction de liens

"Qui influença J.K. Rowling ?"

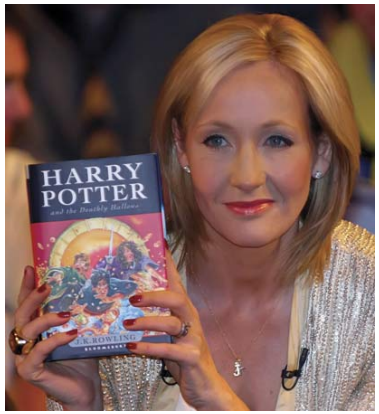
J. K. Rowling `_influenced_by` ?



Prédiction de liens

"Qui influença J.K. Rowling ?"

J. K. Rowling `_influenced_by`



- G. K. Chesterton
- J. R. R. Tolkien
- C. S. Lewis
- Lloyd Alexander
- Terry Pratchett
- Roald Dahl
- Jorge Luis Borges
- Stephen King
- Ian Fleming

Prédiction de liens

"De quel genre est le film WALL-E ?"

WALL-E `_has_genre` ?

Prédiction de liens

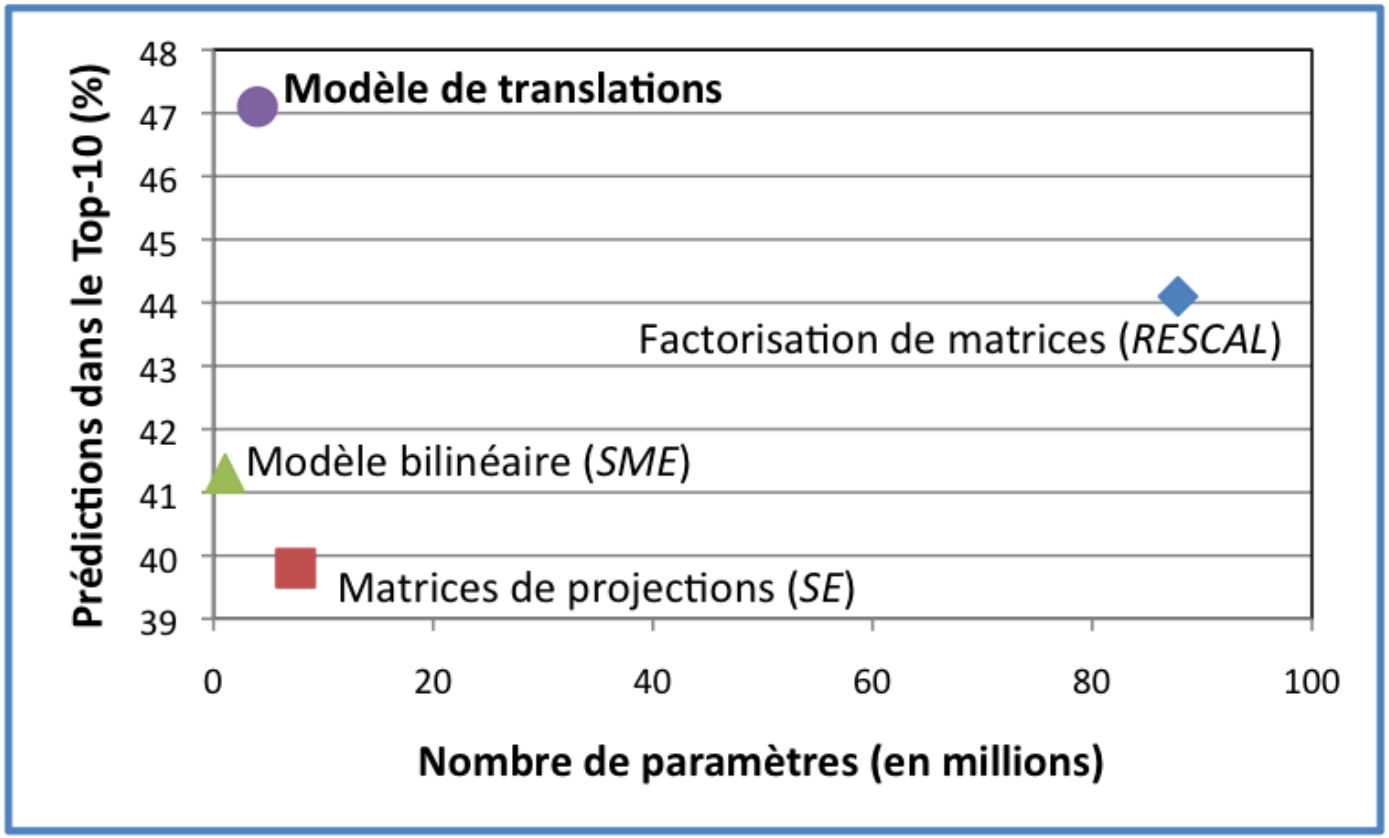
"De quel genre est le film WALL-E ?"

WALL-E `_has_genre`

- Animation
- Computer animation
- Comedy film
- Adventure film
- Science Fiction
- Fantasy
- Stop motion
- Satire
- Drama

Prédiction de liens

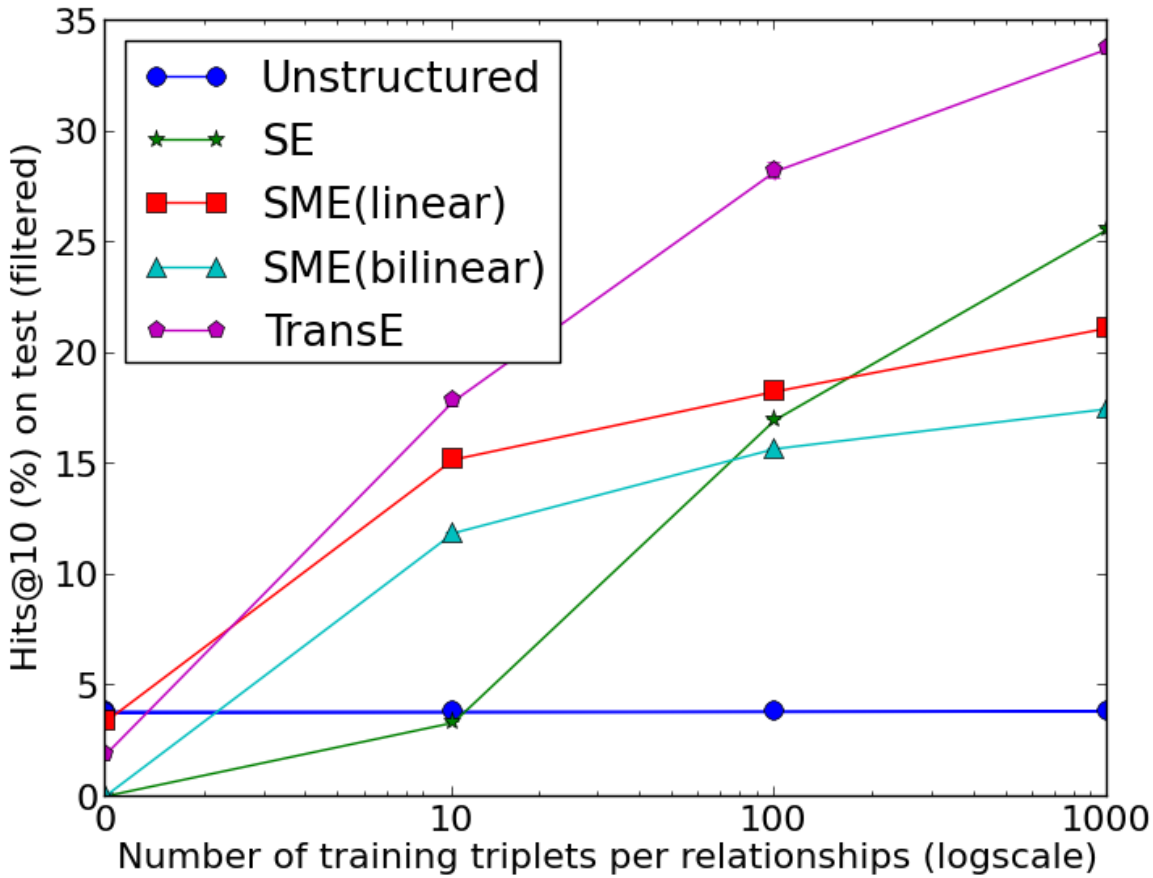
Sur Freebase15k :



Sur Freebase1M, notre modèle prédit 34% dans le Top-10.

Intégrer de nouvelles données

Apprendre les représentations de 40 types de relation inconnus.



Menu

Extraction d'information

- Extraction d'information : texte brut → format structuré.
- Très important de nos jours !
- **But :** compléter des BCs à partir du texte.

- **Notre travail :** utiliser notre approche pour modéliser la BC et améliorer l'extraction.

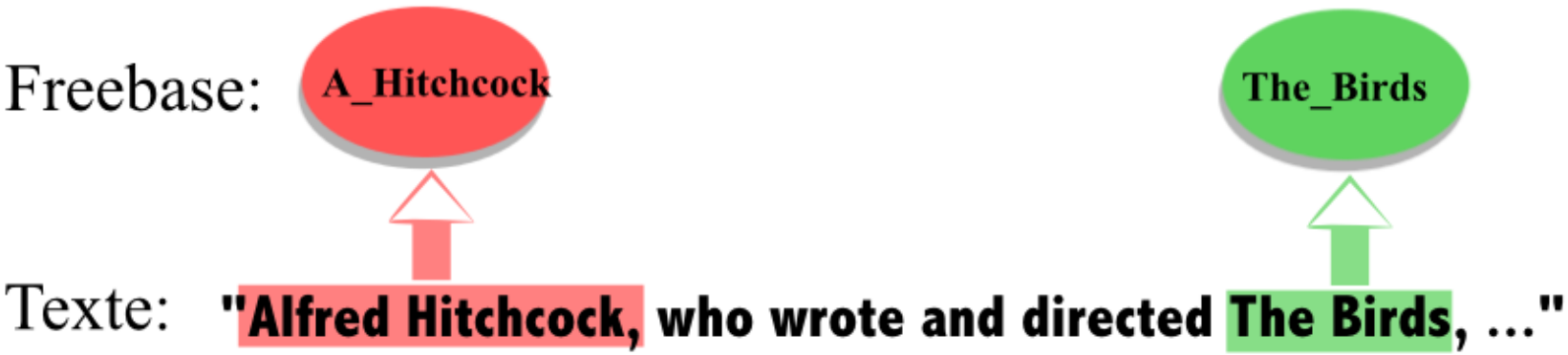
Le problème d'extraction de relations

Etant donnée une phrase.

Texte: **"Alfred Hitchcock, who wrote and directed The Birds, ..."**

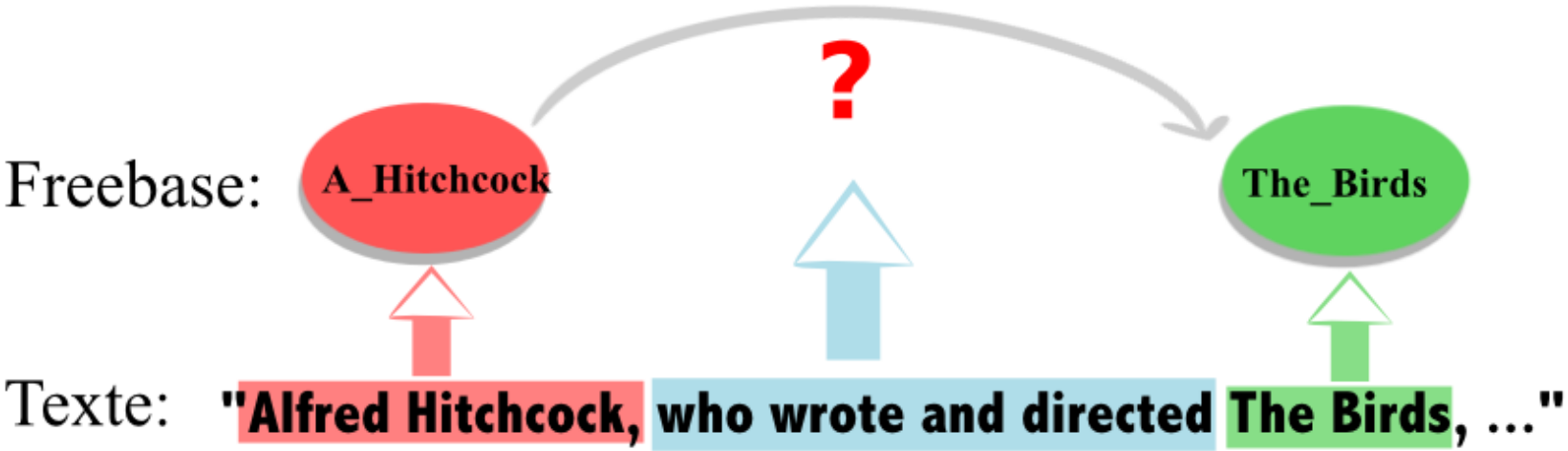
Le problème d'extraction de relations

On considère que les entités ont déjà été détectées.



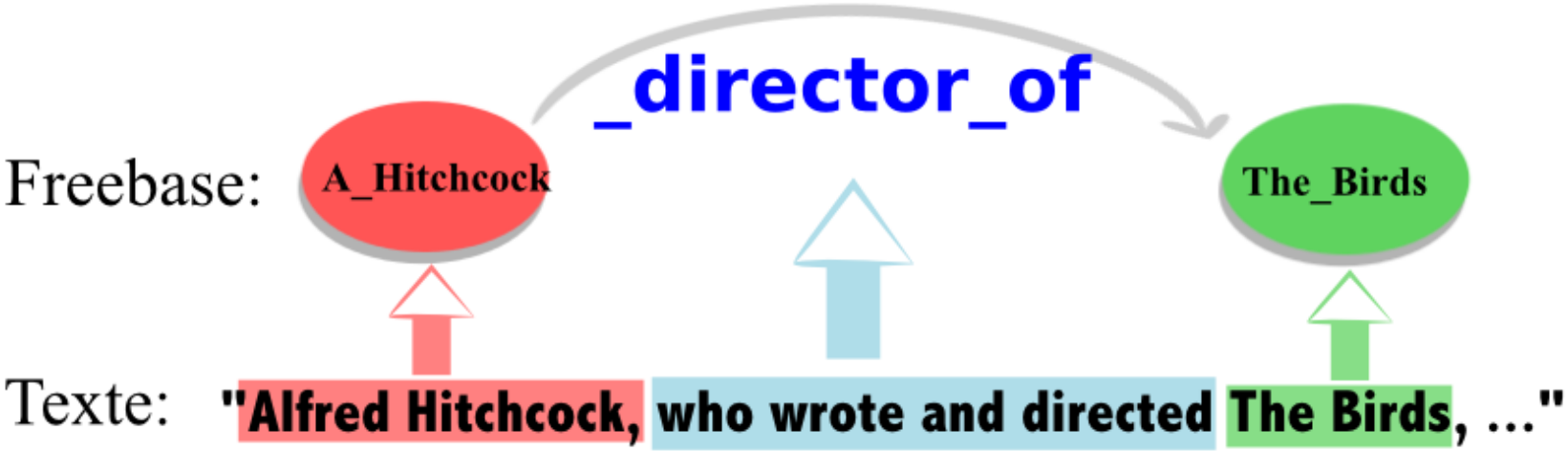
Le problème d'extraction de relations

Le but est d'identifier si une relation existe entre elles.



Le problème d'extraction de relations

Et quel est son type.



Utiliser la BC en plus du texte

- **Méthode classique :** un classifieur est entraîné pour prédire la relation sachant le texte *txt* et les entités *sub* et *obj* :

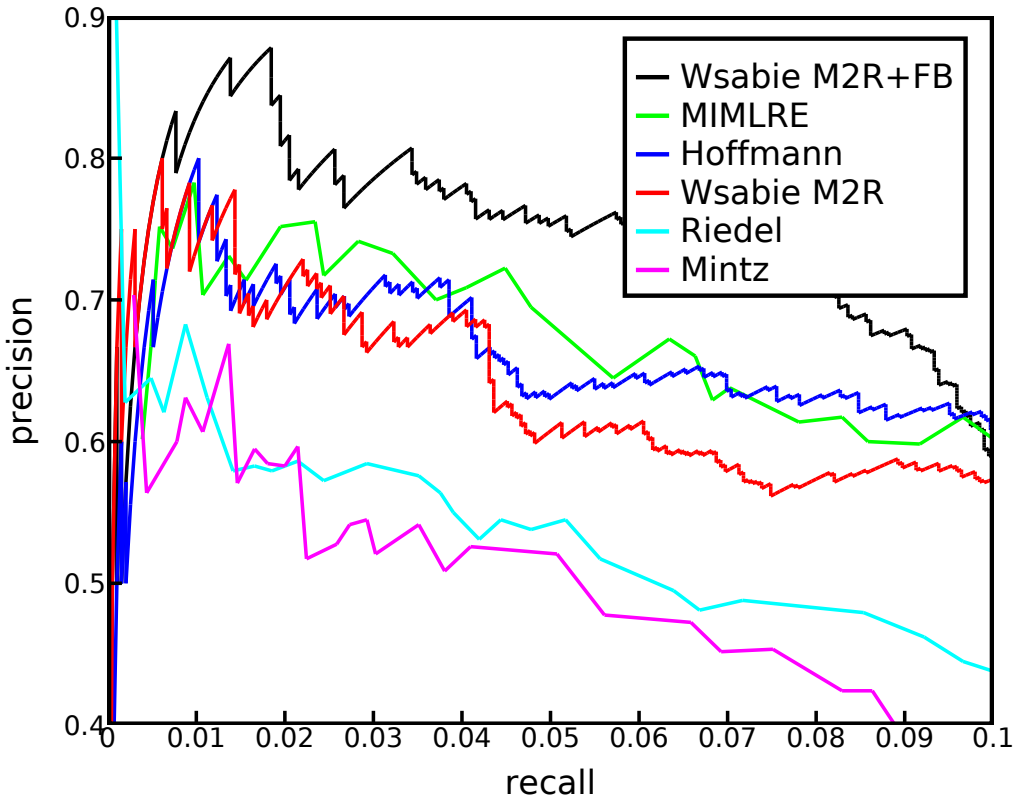
$$r(txt, sub, obj) = \arg\max_{rel'} S_{txt2rel}(txt, rel')$$

- **Idée :** extraire des relations en utilisant le texte + les connaissance existantes (= la BC courante).
- Notre modèle de la BC force la rel. extraite à s'accorder avec :

$$r(txt, sub, obj) = \arg\max_{rel'} (S_{txt2rel}(txt, rel') - d_{BC}(sub, rel', obj))$$

Expériences sur NYT+Freebase

On apprend sur des articles du New York Times et sur Freebase.



Courbe de rappel/précision en prédiction de relations.

Menu

Désambiguïsation dans un cadre particulier

Désambiguïsation → connecter le texte à la BC WordNet.

Vers un système d'interprétation du texte brut :

``A musical score accompanies a television program ."

↓ ***Semantic Role Labeling***

(``A musical score", ``accompanies", ``a television program")

↓ ***Preprocessing (POS, Chunking, ...)***

((_musical_JJ score_NN), _accompany_VB , _television_program_NN)

↓ ***Word-sense Disambiguation***

((_musical_JJ_1 score_NN_2), _accompany_VB_1, _television_program_NN_1)

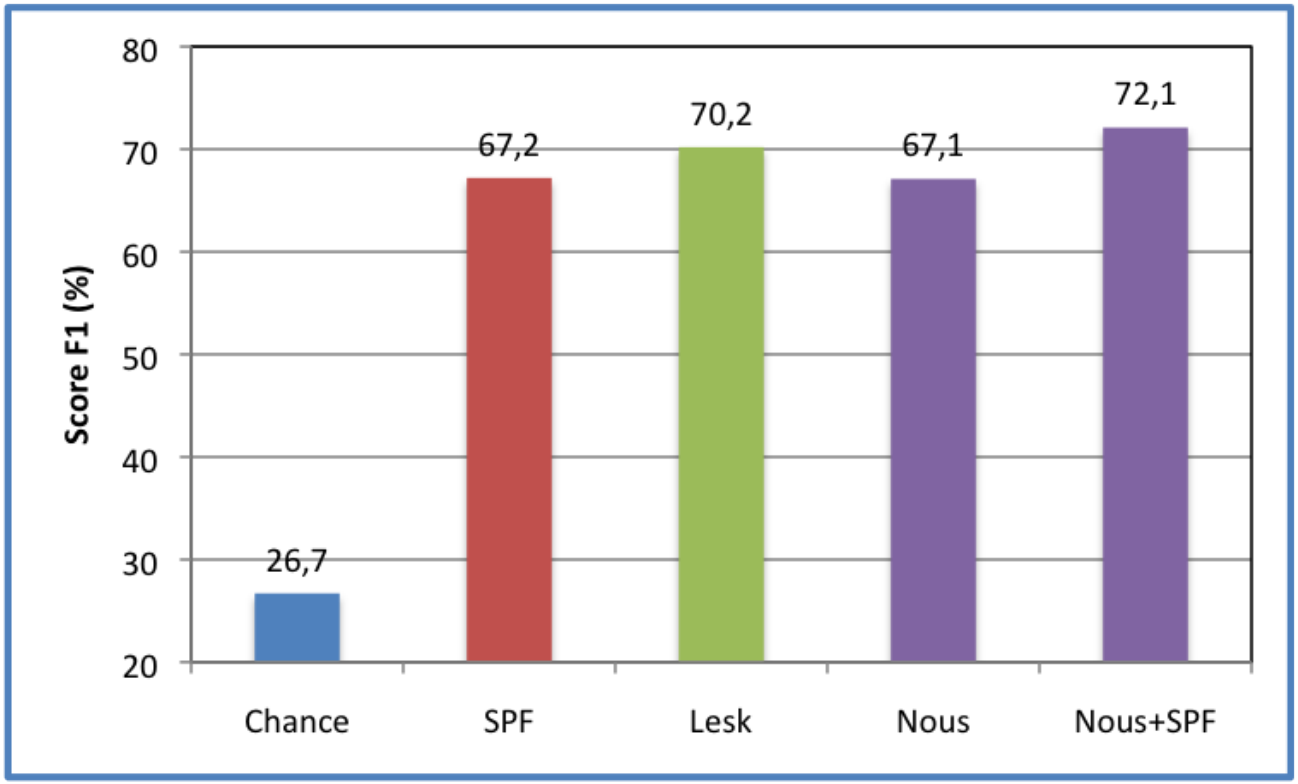
Modéliser texte et BC ensemble

- Le texte est converti en relations (*sub*,*rel*,*obj*).
- On apprend un vecteur pour tout symbole : mots, entités et types de relations de WordNet.
- Le système peut étiqueter 37,141 mots avec 40,943 sens.

	Ex. app.	Ex. test	Etiqueté ?	Symbole
WordNet	146,442	5,000	Non	sens
Wikipedia	2,146,131	10,000	Non	mots
ConceptNet	11,332	0	Non	mots
Ext. WordNet	42,957	5,000	Oui	mots+sens
Unamb. Wikip.	981,841	0	Oui	mots+sens
TOTAL	3,328,703	20,000	-	-

Résultats expérimentaux

Score F1 sur 5,000 phrases de test à désambiguïser.



Compléter WordNet

On crée des similarités qui dépassent WordNet.

"Qu'attaque une armée ?"

army_NN_1 `attack_VB_1` ?

Compléter WordNet

On crée des similarités qui dépassent WordNet.

"Qu'attaque une armée ?"

army_NN_1	`attack_VB_1`	troop_NN_4
		armed_service_NN_1
		ship_NN_1
		territory_NN_1
		military_unit_NN_1

Compléter WordNet

On crée des similarités qui dépassent WordNet.

"Qu'est-ce qui gagne de l'argent ?"

? `earn_VB_1` money_NN_1

Compléter WordNet

On crée des similarités qui dépassent WordNet.

"Qu'est-ce qui gagne de l'argent ?"

person_NN_1 `earn_VB_1` money_NN_1
business_firm_NN_1
family_NN_1
payoff_NN_3
card_game_NN_1

Menu

Encoder les BC dans des espaces vectoriels

- Les BC sont riches mais demandent de l'attention.
- Apprendre à les projeter dans des espaces vectoriels :
  - Facilite la visualisation ;
  - Permet la prédiction de liens (aves/sans données externes) ;
  - Plus simple à utiliser dans d'autres systèmes ;
  - Format compact.
- Permet la valorisation de ces données de BC.

Défis

Ce n'est que le début :

- Arriver à raisonner : combiner logique, déduction.
- Evaluer la fiabilité des prédictions.
- Résumer des BC.
- Fusionner des BC.
- Connecter texte et BC : interactions réciproques.
- Mieux maîtriser l'optimisation.
- Raffiner le modèle de translation.
- etc.

Fin

Beaucoup de code/données sont disponibles depuis ma page.

Merci !

antoine.bordes@hds.utc.fr
http://www.hds.utc.fr/~bordesan

Bibliographie

1. Learning Structured Embeddings of Knowledge Bases.
   A. Bordes, J. Weston, R. Collobert & Y. Bengio. *AAAI, 2011*.

2. Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing.
   A. Bordes, X. Glorot, J. Weston & Y. Bengio. *AISTATS, 2012*.

3. A Latent Factor Model for Highly Multi-relational Data.
   R. Jenatton, N. Le Roux, A. Bordes & G. Obozinski. *NIPS, 2012*.

4. A Semantic Matching Energy Function for Learning with Multi-relational Data.
   A. Bordes, X. Glorot, J. Weston & Y. Bengio. *MLj, 2013.*

5. Irreflexive and Hierarchical Relations as Translations.
   A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston & O. Yakhnenko.
   ICML Workshop on Structured Learning, 2013