Peuplement d'ontologie à partir de petites annonces

Bilan du stage de Quentin Leroy et avancées sur la problématique

Céline Alec

Équipe CoDaG, GREYC, Université Caen-Normandie

18 février 2021

Stage de Quentin Leroy

- Co-encadrement GREYC-LITIS avec Jean-Philippe Kotowicz
- Master 2 DOP (Décision et Optimisation) Université de Caen
- Stage de 6 mois de mars à août 2020 (dont confinement)

Problématique du stage

Peupler une ontologie du domaine des petites annonces immobilières à partir des contenus textuels de petites annonces.
$\Rightarrow$ Arriver à représenter les textes sous une forme structurée

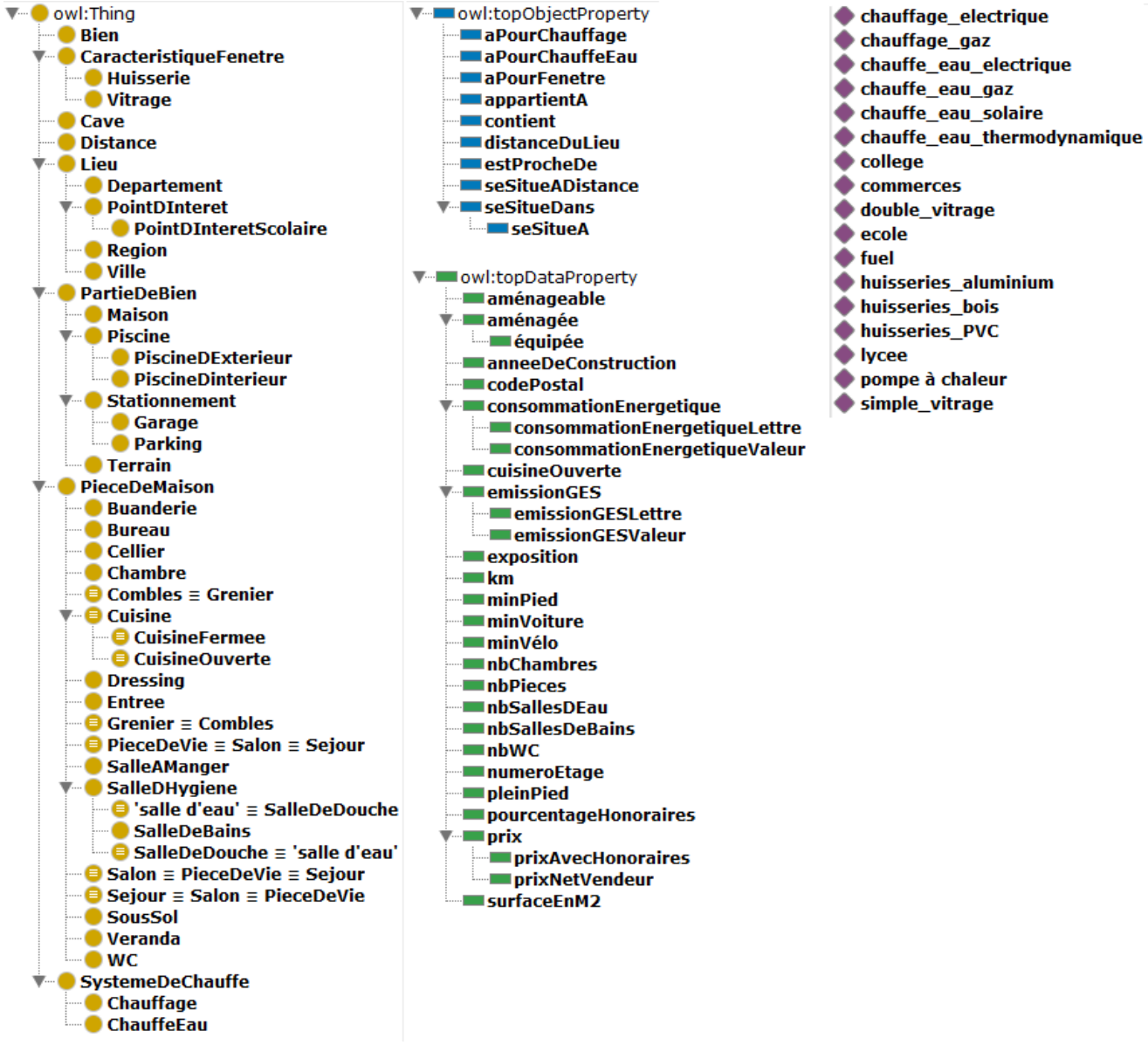
« Maison, à 20 min de Caen, de 125 m^2 avec un terrain de 500 m^2. » :

- <bien1, rdf:type, Bien>
- <bien1, contient, maison1>
- <maison1, rdf:type, Maison>
- <bien1, seSitueADistance, distance1>
- <distance1, rdf:type, Distance>
- <distance1, minVoiture, 20>
- <distance1, distanceDuLieu, Caen>
- <maison1, surface, 125>
- <bien1, contient, terrain1>
- <terrain1, rdf:type, Terrain>
- <terrain1, surface, 500>

Objectifs

- Création d'un corpus d'annonces immobilières : stagiaire L3
- Création de l'ontologie : focus sur les ventes de maisons
- État de l'art
- Proposition d'un début de réflexion sur un processus de peuplement automatique, avec un objectif de généricité : même processus pour divers domaines de petites annonces

Création de l'ontologie



État de l'art

Une étude récente sur le peuplement d'ontologie [Lubani et al., 2019].

Les approches se basent sur :

- des règles lexico-syntaxiques
- de l'apprentissage automatique
- des systèmes hybrides

Utiliser de l'apprentissage automatique nécessite de disposer en amont d'une quantité suffisante de phrases et de leur correspondance ontologie.

État de l'art

Diverses approches exploitent des règles lexico-syntaxiques. Quelques exemples sont donnés ici.

- [Hearst, 1992] : patrons lexico-syntaxiques permettant d'identifier des relations d'hyponymie (« such as », « is a »). Des versions plus évoluées, permettent d'assister un expert pour identifier les patrons.
- [Alani et al., 2003] : peuplement d'ontologie à partir du web (domaine des artistes). Ne peuple que des assertions. Se base sur le verbe présent dans la phrase entre 2 instances de l'ontologie
- [Makki et al., 2009] : se focalise aussi sur les verbes pour peupler des assertions (semi-automatique et indépendant du domaine)
- [Faria et al., 2014] : cherche à instancier un concept ainsi que les relations qui le concernent. Se base sur les noms propres pour instancier les concepts et sur les verbes pour les assertions.
- [Buitelaar et al., 2006, Weber and Buitelaar, 2006] : utilisent des données structurées en plus des textes

Positionnement

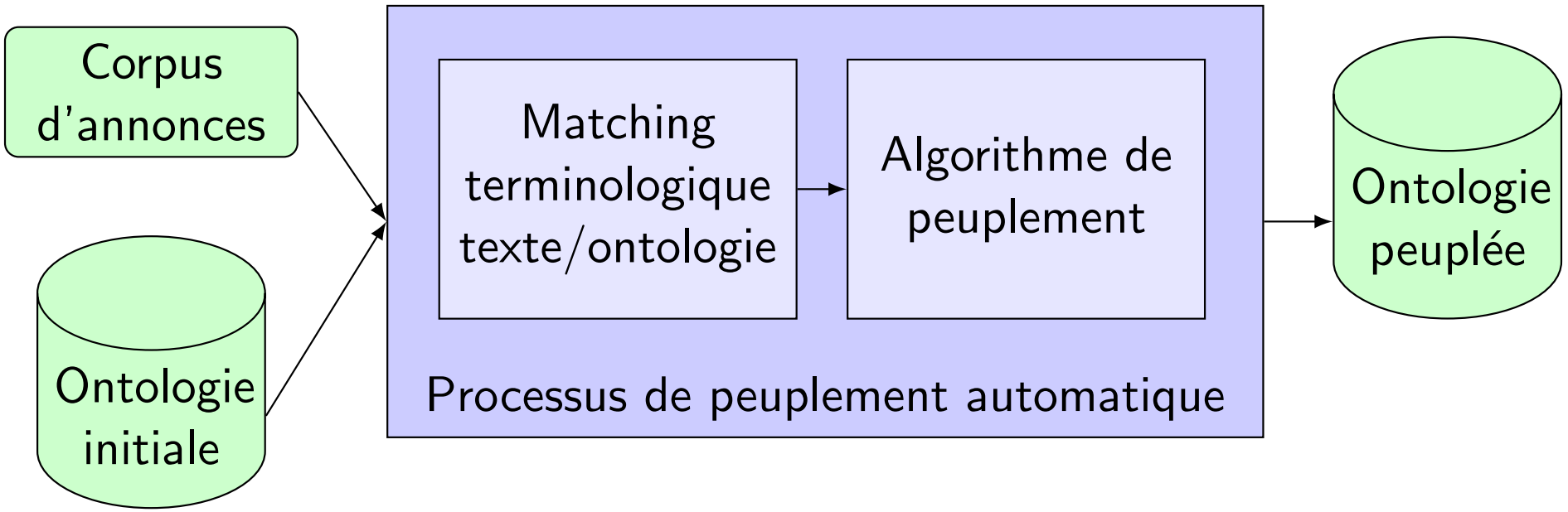
Schéma standard des approches actuelles :

- « sujet verbe complément » ⇒ <sujet> <prédicat> <objet>
- sujets et compléments = entités nommées
- verbes fortement caractéristiques d'une relation (ex : « est marié à »)

Notre contexte :

- pas (ou peu) d'entités nommées
- sujet = le bien à vendre ou quelque chose lié à ce bien
- verbes peu parlants (« a », « possède », « contient », etc.), voire pas de verbe
- propriétés typées à prendre en compte (surfaces, prix, diagnostic énergétique, etc.)
- propriétés non binaires (un bien immobilier se situe à une certaine distance d'un point d'intérêt)

Proposition d'un processus de peuplement



Matching terminologique

Lemmatisation pour prendre en compte les variations morphologiques

Exemples

- chambres ⇒ chambre
- salles de bain ⇒ salle de bain

Prise en compte de la casse (camelCase et snake_case) dans l'ontologie :

Exemples

- cuisineOuverte ⇒ cuisine ouverte
- double_vitrage ⇒ double vitrage

Algorithme de peuplement

Déroulement d'un exemple

« Maison, à 20 min de Caen, de 125 m^2 avec un terrain de 500 m^2. »

On veut obtenir :

- <bien1, rdf:type, Bien>
- <bien1, contient, maison1>
- <maison1, rdf:type, Maison>
- <bien1, seSitueADistance, distance1>
- <distance1, rdf:type, Distance>
- <distance1, minVoiture, 20>
- <distance1, distanceDuLieu, Caen>
- <maison1, surface, 125>
- <bien1, contient, terrain1>
- <terrain1, rdf:type, Terrain>
- <terrain1, surface, 500>

Algorithme de peuplement

Déroulement d'un exemple

« Maison, à 20 min de Caen, de 125 m^2 avec un terrain de 500 m^2. »

Initialisation :
<bien1, rdf:type, Bien>

Matchings :

- classe `Maison`
- propriété typée `minVoiture` avec valeur 20
- individu `Caen` de type `Ville`
- propriété `surface` avec valeur 125
- classe `Terrain`
- propriété typée `surface` avec valeur 500

Algorithme de peuplement

« Maison, à 20 min de Caen, de 125 m^2 avec un terrain de 500 m^2. »

Analyse des matchings

Classes `Maison` - `Terrain` :

- `<maison1, rdf:type, Maison>`
- `<terrain1, rdf:type, Terrain>`

Individu `Caen` de type `Ville`
Propriété `surface` avec valeur 125 et 500 (domaine : `PartieDeBien` ou `PieceDeMaison`) :

- `<maison1, surface, 125>`
- `<terrain1, surface, 500>`

Propriété `minVoiture` avec valeur 20 (domaine : `Distance`) :

- `<distance1, minVoiture, 20>`
- `<distance1, rdf:type, Distance>`

Algorithme de peuplement

Analyse du texte

« Maison, à 20 min de Caen, de 125 m^2 avec un terrain de 500 m^2. »

Exploiter les propriétés de l'ontologie et l'ordre des instances obtenues via matchings :
`maison1` - `distance1` - `Caen` - ∅ - `terrain1` - ∅

- `<distance1, distanceDuLieu, Caen>`

Algorithme de peuplement

Analyse hors texte basée sur les connaissances de l'ontologie

On cherche à ce que les individus soient raccrochés les uns aux autres.

- <bien1, contient, maison1>
- <bien1, contient, terrain1>
- <bien1, seSitueADistance, distance1>

Algorithme de peuplement : idées et questionnement

- Si un même mot matche avec plusieurs éléments ontologiques ?
- Choix de la propriété à considérer en cas de multi-candidats ?
- Axiomes ontologiques pas forcément évidents à traiter (domaine/co-domaine inconnu mais déductions possibles)
- Analyse du texte : essayer de raccorder si dans la même phrase ? Trouver un juste milieu pour obtenir des raccords mais ne pas chercher à tout prix à tout raccorder ?
- Analyse hors texte : Idem, ne pas raccorder à tout prix. Choix si plusieurs individus raccordables possibles ?
- Quand exploiter un individu déjà créé et quand en créer un nouveau ?

Conclusion

- Une problématique novatrice par rapport à l'état de l'art
- Une première version d'ontologie proposée
- Un début de réflexion sur une approche de peuplement automatique

Travaux futurs :

- Implémenter une première version de l'algorithme
- Observer les cas complexes et envisager des améliorations du processus de peuplement
- Expérimenter en comparant avec un peuplement manuel
- Tenter l'expérience avec un autre domaine des petites annonces

Merci pour votre attention

Questions ?

Références

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Lewis, P. H., and Shadbolt, N. R. (2003).
Automatic ontology-based knowledge extraction and tailored biography generation from the web.
IEEE Intell Syst, pages 14–21.

Buitelaar, P., Cimiano, P., Racioppa, S., and Siegel, M. (2006).
Ontology-based information extraction with SOBA.
In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy. European Language Resources Association (ELRA).

Faria, C., Serra, I., and Girardi, R. (2014).
A domain-independent process for automatic ontology population from text.
Science of Computer Programming, 95 :26 – 43. Special Issue on Systems Development by Means of Semantic Technologies.

Hearst, M. A. (1992).
Automatic acquisition of hyponyms from large text corpora.
In COLING 1992 Volume 2 : The 15th International Conference on Computational Linguistics.

Lubani, M., Noah, S. A. M., and Mahmud, R. (2019).
Ontology population : Approaches and design aspects.
Journal of Information Science, 45 :502 – 515.

Makki, J., Alquier, A.-M., and Prince, V. (2009).
Ontology population via nlp techniques in risk management.
Int J Hum Soc Sci, pages 212–217.

Weber, N. and Buitelaar, P. (2006).
Web-based ontology learning with isolde.
In Workshop on web content mining with human language at the international semantic web conference.