Apprentissage de structures séquentielles pour l'extraction d'entités et de relations dans des textes d'appels d'offres

Oussama Ahmia*, Nicolas Béchet*, Pierre-François Marteau*,

* IRISA, Université Bretagne Sud, Rue Yves mainguy – BP 573 56000 VANNES cedex
nom.prénom@irisa.fr, http://www-expression.irisa.fr

Résumé. Dans cet article nous présentons une étude exploitant des méthodes d'apprentissage automatique de structures séquentielles pour extraire des relations sémantiques dans des textes issus de bases d'appels d'offres. L'une des relations que nous considérons concerne l'emprise d'un projet d'aménagement, caractérisée par une association entre les concepts qui définissent les infrastructures (bâtiments) et les concepts qui définissent leur(s) surface(s) d'implantation. L'étude propose une analyse comparée d'approches à base de champs conditionnels aléatoires (CRF), de CRF d'ordre supérieur (H-CRF), de CRF semi-Markoviens, Modèles de Markov cachés (HMM) et de perceptrons structurés.

1 Introduction

L'identification des projets d'aménagement futurs permet de représenter la ville de demain, et pour cela, il est nécessaire de disposer en temps opportun de données sur la nature du projet d'aménagement, caractérisée en particulier par sa superficie et sa destination principale (typologie de bâtiment). Les appels d'offres relatifs aux marchés publics constituent ainsi une source importante d'informations. Sur la base d'un jeu de données collecté à partir du Bulletin Officiel des Annonces des Marchés Publics (BOAMP), notre étude concerne la détection de la typologie des bâtiments référencés dans une annonce et leurs superficies respectives.

L'extraction d'information que nous ciblons ici, consiste à extraire automatiquement des données structurées telles que des entités (nommées), des relations entre entités, ou encore des attributs décrivant des entités, à partir de sources non structurées (Sarawagi, 2008).

Dans le but d'extraire la surface d'un bâtiment dans des textes d'appels d'offres en français, nous proposons une analyse comparative d'un ensemble de modèles statistiques dont les Champs Aléatoires Conditionnels (CRF).

Différentes caractéristiques issues des textes sont exploitées : des caractéristiques grammaticales et sémantiques pour la caractérisations des entrées lexicales, et des ensembles de caractéristiques à *longue portée* (décrivant le contexte) qui ont démontré leur efficacité dans le domaine d'étiquetage des séquences symboliques (Li et al., 2011).

2 État de l'art

La littérature fait état de nombreuses méthodes pour la détection d'entités et l'extraction de relations entre les entités. Ces méthodes sont regroupées en général en deux catégories : les approches supervisées et les approches non supervisées.

Les approches non-supervisés se basent souvent sur des caractéristiques contextuelles. La sémantique distributionnelle, introduite par Z.S Harris (Harris, 1954), considère que deux entités qui co-occurrent fréquemment dans des contextes similaires ont tendance à partager un même sens. D. Ravichandran utilise un *bootstrap* (Ravichandran et Hovy, 2002) afin d'apprendre des motifs de surface dans le but d'extraire des relations binaires à partir du Web. Plus récemment (Min et al., 2012) ont développé une méthode d'extraction de relations appliquée à des données à grande échelle qui permet de gérer de manière très générale les problèmes de polysémie et de synonymie, sources d'ambiguïtés sémantiques difficiles à surmonter pour l'extraction de certaines classes de relation.

Généralement, les approches supervisées sont subdivisées en deux sous-catégories : les méthodes basées sur des noyaux (kernels) ou les méthodes basées sur les caractéristiques (features). Le principe des méthodes basées sur les caractéristiques est d'extraire un ensemble de variables syntaxiques et sémantiques pour l'apprentissage statistique (Miller et al., 2000). Plusieurs techniques ont été utilisées dans le cadre cette approche : l'utilisation de méthodes de classification de structures tels que les CRF linéaires (Banko et al., 2008) et plus récemment une approche semi-supervisée ("Distant supervised") a été proposée afin d'extraire des relations dans des grands volumes de texte ((Angeli et al., 2014)). La méthodes basées sur les noyaux aborde le problème d'extraction de relations sous la forme d'une classification de paire d'entités (mises en relation binaire) (Zelenko et al., 2003). Le modèle de classification utilisé exploite un noyau basé sur l'arbre syntaxique d'une phrase.

Qian et al (Qian et al., 2008) ont développé une approche dynamique pour déterminer les sous arbres qui encodent potentiellement des relations, en exploitant des noyaux d'arbres syntaxiques. Plus récemment Zhou et al (Zhou et al., 2010) ont développé une nouvelle approche qui utilise des informations syntaxiques et sémantiques enrichies pour développer un noyau convolutionnel sensible au contexte. Ce type de noyau permet d'extraire des sous arbres modélisant des relations qui tiennent compte du contexte.

Disposant d'un jeu de données textuelles étiqueté issu du BOAMP, nous proposons dans cet article une étude comparative de méthodes d'apprentissage supervisé de structures séquentielles pour résoudre le problème d'extraction de relations proposé.

3 Méthodologie

3.1 Approche utilisée

L'extraction d'une relation sémantique se divise en deux étape : (1) la détection d'entités nommées (surface, bâtiment ou autre) et (2) la recherche de relations de type **"est surface de"** entre les entités. Ces relations permettent de lier un bâtiment à une surface (p. ex. *une* ***piscine municipale*** de ***1000 m***^{2}) mais aussi de lier plusieurs bâtiments à une seule surface (p. ex. *Environ* ***2500 m2*** *de plancher, comprenant des* ***bureaux****, des* ***salles de lecture***, etc).

L'approche proposée consiste à effectuer simultanément l'extraction d'entités nommées (surface et bâtiment) et la détection des relations potentiellement existantes entre ces entités. Cette approche peut ainsi être appréhendée comme un problème d'**étiquetage de séquences**, avec pour étiquettes possibles : **O** : les éléments qui ne relèvent pas des catégories bâtiment ou surface, etc. ; **LIEN_SURFACE/bâtiment** : bâtiments de la relation **"est surface de"** ; **O/bâtiment** : bâtiments hors de la relation **"est surface de"** ; **O/surface** : surfaces hors de la relation **"est surface de"** ; **LIEN_SURFACE/surface** : surfaces de la relation **"est surface de"** ; **LIEN_SURFACE** : items de la relation **"est surface de"**, mais qui ne sont ni surface ni bâtiment (cette étiquette permet de lier plus aisément les entités entre elles dans le cas où il existe plusieurs surfaces et bâtiments mis en relation dans une même séquence).

La figure 1 permet de visualiser le résultat escompté par notre modèle d'étiquetage :

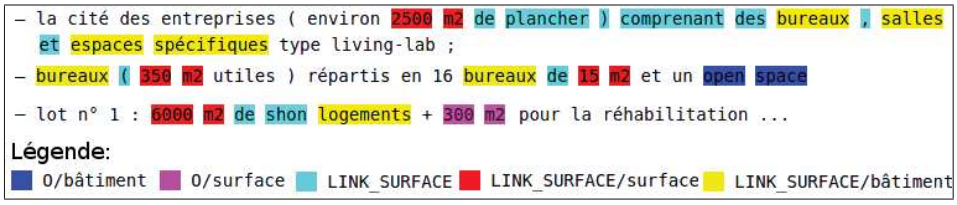

FIG. 1 – *Illustration des étiquettes utilisées*

A partir de l'étiquetage de la première phrase dans la figure 1, on peut facilement en déduire la relation suivante : *"2 500 m2"* ***est surface de*** *"bureaux + salles + espaces spécifiques"*, et pour cela il suffit d'extraire une suite d'éléments (items) qui ont pour étiquettes : **LIEN_SURFACE/bâtiment**, **LIEN_SURFACE/surface** ou **LIEN_SURFACE**.

3.2 Caractéristiques utilisées

Nous avons utilisé différents types de caractéristiques pour représenter les contenus des textes. Des caractéristiques locales relatives a un item de la séquence (entrée lexicale) :

word.lower (la forme du mot en minuscule), **istitle** (indique si le mot commence par une majuscule), **lemma** (le lemme du mot), **POS** ("Part Of Speech", la classe grammaticale du mot), **type** (indique si le mot est **bâtiment**, **surface** ou **0** autrement). Et des caractéristiques à longue portée qui modélisent le contexte d'occurrence des mots qui permet en général de réduire l'ambiguïté sémantiques des termes, notamment efficaces pour l'étiquetage de séquence (Li et al., 2011). Soit x'_i le vecteur des caractéristiques associé à un terme w_i une fois ajoutées les caractéristiques à longue portée. x'_i est construit à partir des caractéristiques locales x_i, de la manière suivante : $x'_i = \langle x_{i-j}, ..., x_{i+j} \rangle$, où j représente la taille du contexte. Par exemple, pour $j = 2$, $x'_4 = \langle x_2, x_3, x_4, x_5, x_6 \rangle$.

3.3 Données utilisées

Nous avons indexé les données collectées sur le site BOAMP en utilisant un moteur de recherche (Lucene) afin de filtrer les annonces les plus concerner la construction de nouveaux bâtiments. La description de chaque annonce est ensuite découpée en phrases puis, à l'aide

d'expressions régulières dédiées, les présences de terminologies décrivant une surface sont détectées dans la phrase. Les phrases qui contiennent des surfaces sont ensuite segmentées en mots, puis les caractéristiques locales et contextuelles sont extraites (cf. section 3.1) pour obtenir un ensemble de 2000 séquences à étiqueter [1].

3.4 Résultats expérimentaux

Différents algorithmes ont été évalués sur les jeux de données présentés. Divers types de champs conditionnels aléatoires (CRF) : avec des caractéristiques à longue portée en faisant varier la taille du contexte, des CRF d'ordre supérieur (H-CRF) et des CRF semi-Markoviens en faisant varier l'ordre pour ces deux derniers modèles. Nous avons également évalué des modèles de Markov cachés (HMM) et un perceptron structuré. Nous avons utilisé la méthode Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) (Zhu et al., 1997), afin d'optimiser les poids des CRF. Par ailleurs, 80% du jeu de données présenté en section 3.3 est utilisé en tant que données d'apprentissage et 20% comme données de test, en procédant à une validation croisée (k=5). Les score obtenus sont synthétisés dans le tableau 1).

TAB. 1 – *Résultats obtenus avec les différents algorithmes.*

	F1-mesure	exactitude empirique	Écart type
CRF contexte(3)	**0.932**	**76.04%**	1.10%
CRF contexte(2)	0.926	74.65%	1.61%
CRF linéaire	0.897	61.75%	2.51%
CRF semi-Markoviens	0.897	67.74%	1.69%
HCRF ordre(3)	0.877	66.36%	1.36%
HCRF ordre(2)	0.882	63.59%	1.38%
Perceptron structuré	0.897	64.52%	1.84%
Automate (Regex)	0.855	66.89%	0%
HMM	0.667	15.21%	0.28%

3.5 Discussion

Le tableau 1 montre que les CRF avec des variables longues portées obtiennent le meilleur score de **76.04%** pour l'exactitude empirique sur les données de test. Ce résultat montre que l'ajout de caractéristiques décrivant le contexte est efficace et améliore progressivement les résultats. Une taille de contexte supérieure à 3 n'améliore cependant plus les résultats. Un accroissement de la complexité du modèle et le manque de données d'apprentissage peut ici conduire à un sur-apprentissage (*over fitting*). Les CRF semi-Markoviens produisent des résultats inférieurs, avec un score de **67.74%** pour l'exactitude empirique. Finalement, l'algorithme des HMM est le moins performant avec un score de **15.21%** pour l'exactitude empirique.

Le tableau 2, présente les résultats par étiquette pour le meilleur modèle (**CRF contexte (3)**).

1. la base finale étiquetée sera prochainement mise à disposition de la communauté

TAB. 2 – *Tableau des scores du CRF contexte (3) par étiquette.*

	précision	rappel	f1-mesure	support
O	0.954	0.943	0.948	3473
LINK_SURFACE	0.891	0.909	0.900	1764
LINK_SURFACE/bâtiment	0.931	0.940	0.935	315
O/bâtiment	0.807	0.800	0.803	115
LINK_SURFACE/surface	0.973	0.973	0.973	524
O/surface	0.767	0.793	0.780	58

A l'issue d'une analyse approfondie des résultats et des erreurs d'étiquetage, il s'est avéré que quelques phrases étiquetées par les experts présentent quelques imprécisions. Le modèle, en désaccord avec la vérité terrain, a proposé un meilleur étiquetage dans la plupart des situation erronées. On estime à 1% le nombre de mauvaises annotations manuelles.

En analysant les poids du modèle CRF à l'issue de l'apprentissage, on peut également noter que la variable de **type** aide pour la classification des concepts surface et bâtiment mais que le CRF ne se base pas uniquement sur cette variable d'ordre sémantique pour établir sa prédiction : comme on peut le vérifier dans le tableau 3, l'étiquette POS :NOM est également discriminante pour identifier le concept de bâtiment.

TAB. 3 – *Extrait des poids associés aux caractéristiques les plus discriminantes (positivement ou négativement), par étiquette, pour le modèle CRF contexte (3).*

Variable	Étiquette	Poids
lemma :m2	LINK_SURFACE/surface	1.087
type :surface	LINK_SURFACE/surface	2.296
type :bâtiment	LINK_SURFACE/bâtiment	4.674
word.lower() :m2	LINK_SURFACE/surface	1.079
POS :NOUN	LINK_SURFACE/bâtiment	1.561
lemma :local	O	-0.81

4 Conclusion

Dans cet article nous avons comparé plusieurs algorithmes de prédiction de structures séquentielles appliqué à un problème d'extraction simultanée d'entités et de relations entre entités à partir de données textuelles. Les résultats montrent que les CRF sont significativement plus adaptés à ce type de tâche. Par ailleurs, l'utilisation de caractéristiques à longues portées décrivant des contextes d'occurrences des caractéristiques lexicales nous permettent d'améliorer globalement les résultats des modèles CRF.

Références

Angeli, G., J. Tibshirani, J. Wu, et C. D. Manning (2014). Combining distant and partial supervision for relation extraction. In *EMNLP*, pp. 1556–1567.

Banko, M., O. Etzioni, et T. Center (2008). The tradeoffs between open and traditional relation extraction. In *ACL*, Volume 8, pp. 28–36.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Li, Y., J. Jiang, H. L. Chieu, et K. M. A. Chai (2011). Extracting relation descriptors with conditional random fields. In *IJCNLP*, pp. 392–400.

Miller, S., H. Fox, L. Ramshaw, et R. Weischedel (2000). A novel use of statistical parsing to extract information from text. In *Proc. of the 1st North American Chapter of the ACL Conference*, NAACL 2000, Stroudsburg, PA, USA, pp. 226–233. ACL.

Min, B., S. Shi, R. Grishman, et C.-Y. Lin (2012). Ensemble semantics for large-scale unsupervised relation extraction. In *Proc. of the 2012 Joint Conference on EMNLP and CoNLL*, EMNLP-CoNLL '12, Stroudsburg, PA, USA, pp. 1027–1037. ACL.

Qian, L., G. Zhou, F. Kong, Q. Zhu, et P. Qian (2008). Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 697–704. ACL.

Ravichandran, D. et E. Hovy (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 41–47. Association for Computational Linguistics.

Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases 1*(3), 261–377.

Zelenko, D., C. Aone, et A. Richardella (2003). Kernel methods for relation extraction. *Journal of machine learning research 3*(Feb), 1083–1106.

Zhou, G., L. Qian, et J. Fan (2010). Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *Information Sciences 180*(8), 1313–1325.

Zhu, C., R. H. Byrd, P. Lu, et J. Nocedal (1997). Algorithm 778 : L-bfgs-b : Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS) 23*(4), 550–560.

Summary

In this article we present a study exploiting machine learning methods for sequential structures extraction dedicated to extract semantic relations in call for tender databases on public facilities projects. One of the relationships we consider concerns the impact of a development project. We characterize it as an association between the concepts that define the infrastructure (buildings) and the concepts that define their implantation, namely surfaces. This sequential structure extraction paradigm is considered as a labeling problem of sequential data. A comparative is carried out exploiting several statistical learning techniques. This study demonstrates the robustness of the CRF model for this kind of task when *long term* characteristics that describe the contexte of occurrence of the labels are taken into account.