Ubiq : une plateforme de collecte, analyse et valorisation des corpus

François-Régis Chaumartin[1]
(1) Proxem, 19 boulevard de Magenta, 75010 Paris
frc@proxem.com

Résumé. Proxem édite Ubiq, une plateforme de collecte de documents et d'analyse sémantique, capable d'extraire des informations pertinentes à partir du contenu de vastes corpus. Les documents analysés sont d'une grande diversité : opinions collectées sur des sites web, emails de réclamation ou de demande d'information, réponse à des questions ouvertes dans des sondages, offres ou demandes d'emploi, etc. La reconnaissance des entités nommées joue un rôle central car c'est un préalable à d'autres traitements sémantiques. La conception d'un module de reconnaissance d'entités nommées nécessite généralement un investissement important en amont, avec une adaptation de domaine. Ubiq propose une approche d'apprentissage faiblement supervisé de l'extraction d'entités nommées qui tient compte du corpus collecté et de ressources externes (Wikipédia). La méthode et l'outillage développés permettent de déterminer à la volée, en interaction avec l'utilisateur, la granularité des types d'entités adaptée à un corpus de texte tout-venant.

Abstract. Proxem publishes Ubiq, a platform for web crawling and semantic analysis, which can extract relevant information from large corpus. Documents are of great variety: reviews crawled from websites, emails about complaints or requests for information, answers to open questions in surveys, employment offers or job applications, etc. Named Entity Recognition plays a key role since it is a prerequisite to further semantic processing. The design of a NER module generally requires a significant upfront investment with some domain adaptation. Ubiq proposes a semi-supervised approach to NER that takes into account the crawled corpus and external resources (Wikipedia). The proposed method and tools allow to get on the fly, with some user interaction, the type granularity of entities suitable for a given corpus.

Mots-clés : entités nommées, désambiguïsation, apprentissage, Wikipédia, catégorisation.
Keywords: named entities, disambiguation, machine learning, Wikipedia, categorization.

1 Objectif : amorcer, à la volée, l'extraction d'entités nommées d'un corpus

La multiplication d'avis de consommateurs sur le web permet aujourd'hui d'effectuer des enquêtes dans divers domaines applicatifs en allant chercher des documents à analyser sur Internet. Ce type d'enquête est généralement mené avec des phases (i) de collecte de documents textuels à partir de sources web, (ii) d'analyse sémantique des contenus et (iii) de présentation de ces informations. La phase d'analyse des contenus débute par des tâches comme le découpage de chaque page web en zones, l'identification de la langue du texte et éventuellement une correction orthographique. L'extraction des entités nommées est réalisée après ces prétraitements ; elle associe un type à chaque instance, dont certaines peuvent être ambiguës (Orange$_{[Fruit]}$ et Orange$_{[Marque\ télécom]}$ par exemple). D'autres traitements sémantiques peuvent suivre : extraction de relations entre entités, analyse d'opinions, classification, priorisation… La qualité de ces traitements dépend directement de celle obtenue lors de la reconnaissance des entités nommées. Sur des projets du domaine de la grande distribution, nous avons atteint une F-mesure de 97% dans la détection des marques et produits ; néanmoins, ce résultat n'a été possible qu'au prix d'un effort manuel significatif ; des semaines de travail ont été nécessaires pour créer les ressources linguistiques et résoudre les principales ambiguïtés présentes. La variété des sujets abordés est le sujet le plus épineux dans ce cadre traité. Lors de nos projets, nous avons été confrontés à la difficulté de constitution des ressources lexicales dans des domaines applicatifs multiples : banque de détail, assurance, automobile, recrutement, télécommunications, mode, bricolage, cosmétique, pétrole, industrie du vin, pathologies médicales… La récurrence de cette problématique nous a poussé à développer différentes approches permettant d'être opérationnel rapidement sur une thématique nouvelle. Ubiq permet la mise en œuvre simultanée de deux approches complémentaires :

– La première a pour objectif la rapidité de mise en œuvre, et consiste à faciliter l'utilisation d'un annotateur générique préexistant comme le système de classification thématique générique décrit dans (Chaumartin, 2013) ou Open Calais : un tel annotateur est prévu pour reconnaître un jeu d'entités prédéfini apparaissant dans des documents d'un certain type (par exemple personnes, lieux et organisations au sein d'articles de presse). Son intérêt est d'être opérationnel immédiatement… sous réserve d'être adapté au corpus à traiter (faible rappel sur un corpus arbitraire).

– La seconde vise à améliorer la finesse d'analyse, en réalisant un annotateur sur mesure, spécifique à un domaine ou à un corpus particulier. Il devient alors possible d'identifier les entités pertinentes d'une façon arbitrairement fine ; dans le domaine du vin, on peut par exemple regrouper toutes les boissons alcoolisées ensemble, ou au contraire choisir de distinguer les vins, bières ou apéritifs cités dans le corpus. Mais l'investissement préalable (pour écrire des règles, annoter manuellement un corpus d'apprentissage ou valider des ressources) est généralement important.

Proxem a innové sur ce dernier point en développant, grâce à la plate-forme Antelope (Chaumartin, 2012), une méthode qui fournit rapidement un annotateur capable d'extraire des entités nommées adapté à tout corpus monolingue (sous réserve qu'il soit relativement homogène). Le système est suffisamment simple à administrer pour être mis en œuvre par un utilisateur sans compétence linguistique forte. Le processus d'amorce d'un tel annotateur repose sur un apprentissage faiblement supervisé qui vise à déterminer à la volée la granularité des types d'entités, avec les interactions suivantes :

1. L'utilisateur constitue un corpus homogène (par exemple par collecte ciblée sur le web).
2. Le système calcule un graphe de thématiques associées à ce corpus, en prenant comme référence les catégories de la Wikipédia dans la langue du corpus.
3. L'utilisateur valide au sein du graphe les catégories proposées qui semblent pertinentes et supprime les autres.
4. Le système restreint les entités nommées candidates aux articles de la Wikipédia rattachés à ces catégories et susceptibles de définir une entité. Ces données alimentent un composant de REN fonctionnant avec des listes d'entités et capable de gérer l'éventuelle homonymie avec des informations de désambiguïsation.
5. Le système effectue une première REN sur le corpus, et compte le nombre effectif d'occurrences de chaque entité candidate. Par ailleurs, l'appartenance d'une entité à une classe et les relations d'hyperonymie entre classes ont été précalculées. Cette double information permet d'afficher à l'utilisateur une arborescence de classes d'entités potentiellement pertinentes dans le contexte du corpus.
6. L'utilisateur amorce une taxonomie ad-hoc en sélectionnant les classes qu'il souhaite reconnaître. Il peut éventuellement en fusionner (pour regrouper des classes Marque, Société et Entreprise, par exemple). Il peut aussi choisir unitairement les entités à conserver ou à exclure.
7. Le système annote à nouveau le corpus, en tenant compte des choix de l'utilisateur. Des post-traitements aident à identifier de nouvelles entités candidates (absentes de la Wikipédia, ou non liées à l'une des catégories choisies lors de la première interaction, ou dont l'hyperonyme n'a pas été précalculé correctement).
8. L'utilisateur valide parmi ces nouvelles entités proposées celles qu'il estime pertinentes.
9. Le système entraîne un composant d'apprentissage (par CRF) sur ce corpus annoté.

2 Résultats préliminaires

Nous avons évalué ce système sur un corpus de 15 000 avis de consommateurs sur l'univers mode & chaussure, le système propose à l'utilisateur les classes d'entités suivantes : 73 entreprises, 69 textiles, 55 photographes, 53 accessoires, 38 types de chaussures. L'utilisateur peut corriger d'éventuelles erreurs et préciser les classes que le système devra réellement reconnaître. La table 1 compare les résultats obtenus en quelques heures par ce système avec ceux d'un analyseur dont la création manuelle a nécessité deux semaines de travail avec une démarche classique. Les résultats expérimentaux sont encourageants ; les entités extraites sont de plus liées aux entrées correspondantes de Wikipédia. Il reste toutefois à améliorer le rappel de cette méthode et à comparer les résultats obtenus sur des jeux de données standard. L'approche proposée est en principe facilement généralisable à de nouvelles langues.

	Analyseur créé manuellement en 10 jours	Analyseur créé avec la méthode décrite ici
Classe Chaussure	71 entités créées manuellement et repérées dans les documents (63000 occurrences)	145 entités proposées par le système dont 72 présentes dans le corpus (36 000 occurrences)
Entités spécifiques	boots (3 878), spartiate (612), sneaker (534)…	babies (431), motarde (245), brogue (61)…
Entités communes	chaussure (11 110), bottine (3 906), tong (1 563), rangers (540), chaussure de sécurité (81)…	

TABLE 1 : Comparaison entre un analyseur créé manuellement et avec la méthode décrite ici

Références

CHAUMARTIN F.-R. (2012). *Antelope, une plate-forme de TAL permettant d'extraire les sens du texte : théorie et applications de l'ISS*. Thèse de doctorat, Université Paris Diderot.

CHAUMARTIN F.-R. (2013). Apprentissage d'une classification thématique générique et cross-langue à partir des catégories de la Wikipédia. Actes de *TALN*, 659–666.