Proxem Studio : la plate-forme d'analyse sémantique qui transforme l'utilisateur métier en *text scientist*

François-Régis Chaumartin [1]
(1) Proxem SAS, 105 rue La Fayette, 75010 Paris, France
frc@proxem.com

RESUME

Proxem édite depuis 2011 une plate-forme d'analyse sémantique multilingue utilisé en entreprise pour de multiples usages : relation clients, ressources humaines, veille stratégique… La version la plus récente du logiciel, lancée en mars 2017, lève le principal goulet d'étranglement des outils classiques de text mining : un utilisateur métier devient enfin autonome pour définir lui-même les ressources linguistiques nécessaires à l'analyse sémantique d'un corpus donné. Une fois le corpus chargé, la plate-forme en extrait une terminologie et organise les termes en regroupements hiérarchisés de proto-concepts ; l'utilisateur n'a plus qu'à valider ces concepts au niveau de granularité qui lui semble pertinent pour constituer un extracteur d'entités nommées de granularité fine, adapté au corpus à traiter, avec un rappel élevé grâce à l'identification automatique des quasi-synonymes. La plate-forme détecte aussi dans ces termes les homonymes potentiels et propose à l'utilisateur des contextes de désambiguïsation, fournissant ainsi une bonne précision.

ABSTRACT

ProxemStudio: the semantic analysis platform that turns the business user into a text scientist. Proxem publishes since 2011 a multilingual semantic analysis platform used by companies for multiple use cases: customer feedback management, human resources, business intelligence. The latest version, launched in March 2017, removes the major bottleneck of conventional text mining tools: a business user finally becomes autonomous to define himself or herself the linguistic resources necessary for the semantic analysis of a given corpus. Once the corpus is uploaded, the platform extracts a terminology and organizes the terms into hierarchical clusters of proto-concepts. The user only has to validate these concepts at the granularity level that seems pertinent to constitute a fine-grained named entities recognizer, perfectly adapted to the corpus, with a high recall thanks to the automatic identification of synonyms. The platform also detects potential homonyms among these terms and provides the user with disambiguation contexts, thus providing a good precision.

MOTS-CLES : entités nommées, catégorisation, désambiguïsation, apprentissage profond.
KEYWORDS: named entities, categorization, disambiguation, deep learning.

1 Evolutions de la plate-forme Proxem en 10 ans

Développée à partir de 2007, la plate-forme Antelope (Chaumartin, 2012) avait pour objectif d'accélérer le travail de l'infolinguiste en lui proposant un cadre de travail simplifié, fourni avec des modules d'analyse prêts à l'emploi. Des algorithmes d'apprentissage ont été intégrés dès 2009 pour la reconnaissance des entités nommées (par CRF), puis en 2010 pour la classification automatique.

Ils ont été ensuite enrichi par un module de catégorisation générique (Chaumartin, 2013) permettant d'associer finement à un texte tout-venant écrit dans une langue donnée, un graphe de catégories de la Wikipédia dans cette langue. D'autres évolutions ont permis de traiter des corpus multilingues.

A partir de 2013, Proxem a commencé à s'approprier les techniques à base de réseaux de neurones, et a généralisé l'approche de type *word embedding* à l'apprentissage simultané de plusieurs langues (Coulmance *et al.*, 2015).

Une réflexion ergonomique a été menée courant 2016 pour déterminer comment proposer à un utilisateur métier une application web simple d'utilisation, intégrant tous ces modules. Cette application, le Proxem Studio, a été lancée en mars 2017. Elle permet à présent à un utilisateur qui n'a pas de compétences particulières en linguistique d'être autonome pour réaliser, de bout en bout, un projet comportant des tâches de web mining, text mining et data mining. Cette application permet de transformer facilement une source web ou un corpus d'entreprise en un nouveau canal de données originales, de procéder à une visualisation riche des données extraites du texte, et d'en tirer une connaissance permettant de prendre des meilleures décisions sur les enjeux métiers.

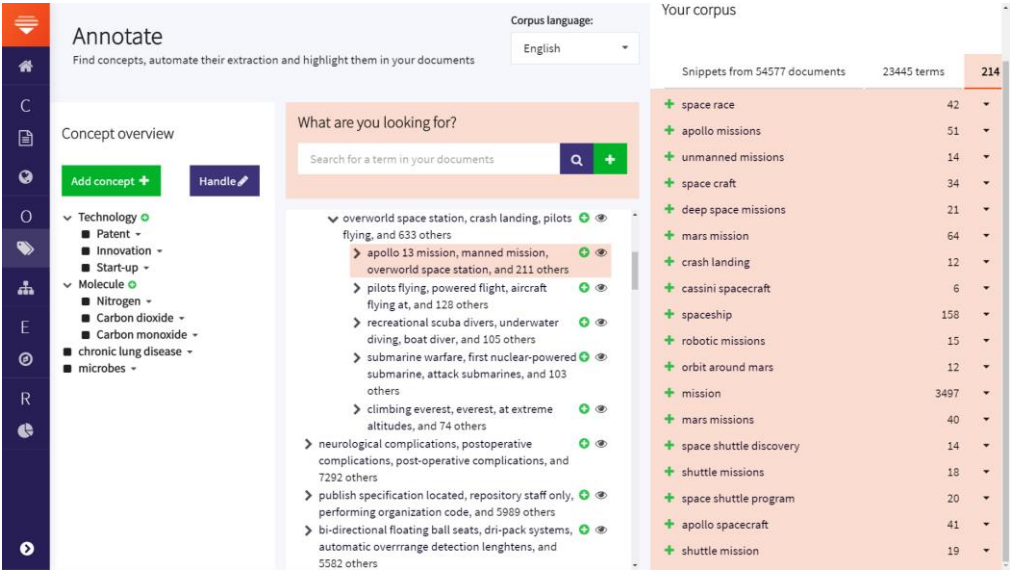

FIGURE 1: Module qui permet à l'utilisateur métier de définir les entités nommées à extraire.

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.

COULMANCE J., MARTY J.-M., WENZEK G., BENHALLOUM A. (2015). Trans-gram, Fast Cross-lingual Word-embeddings. Actes de EMNLP.