Outil Interactif et Évolutif pour l'Extraction d'Information dans des Documents Techniques

Thiziri Belkacem, Charles Teissèdre

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Outil Interactif et Évolutif pour l'Extraction d'Information dans des Documents Techniques

Thiziri Belkacem Charles Teissèdre
Synapse Développement, 7 Boulevard de la gare, 31500 Toulouse, France
thiziri.belkacem@synapse-fr.com; charles.teissedre@synapse-fr.com

RÉSUMÉ

L'accès à l'information dans la documentation technique est une application particulière et complexe du traitement du langage naturel et de la recherche d'information. La difficulté tient aux contraintes propres des langages métier spécialisés et semi-contrôlés. Dans ce document, nous proposons un outil d'accès à l'information dans différents types de documents. Notre solution exploite conjointement la structure organisationnelle des documents et leur contenu informationnel, pour extraire des informations métier dans des différents corpus. Nous proposons un système basé sur des interactions expert-machine dans un cycle d'amélioration continu des modèles d'extraction. Notre approche exploite des modèles d'apprentissage à faible supervision ne nécessitant pas d'expertise en ingénierie des langues. Notre système intègre l'utilisateur dans le processus de qualification de l'information et permet de guider son apprentissage, afin de rendre ses modèles plus performants au fil du temps.

ABSTRACT

Interactive and Evolutive Tool for Information Extraction in Technical Documents

Information access in technical documentation is a particular and complex application of natural language processing and information retrieval. The difficulty lies in the specific constraints of specialized and semi-controlled specialized business languages. In this paper, we propose a tool for accessing information in different types of documents. Our solution jointly exploits the organizational structure of documents and their information content to extract specific pieces of business information from different corpora. We propose a system based on expert-machine interactions in a cycle of continuous improvement of extraction models. Our approach exploits weakly supervised learning models that do not require expertise in language engineering. Our system integrates the user in the information qualification process and allows guiding the user's learning, in order to make the models more efficient over time.

MOTS-CLÉS : Extraction d'Information, Document Technique, Modèle Évolutif..

KEYWORDS: Information Extraction, Technical Document, Evolutive Model..

1 Description

L'accès à l'information dans des corpus de textes de forte technicité est rendu ardue du fait des contraintes propres aux langages métier spécialisés et régulièrement utilisés dans la documentation des industries. Dans cette dernière, les informations et connaissances pertinentes peuvent se présenter

sous différents formats et avoir une certaine régularité ou norme, dépendante du contenu et propre au domaine.

Indépendamment du domaine, différentes applications de traitement du langage naturel et de recherche d'information nécessitent une indexation des séquences informationnelles. Dans des domaines de spécialité, agréger des données d'entraînement permettant de repérer et indexer de telles séquences implique de mobiliser des experts métier en mesure d'analyser et qualifier le contenu des documents. Or la disponibilité des experts - une ressource rare - constitue un frein à la constitution de telles ressources.

Du fait que les modèles de langue génériques sont généralement moins performants dans les domaines spécifiques (Salloum *et al.*, 2020; Torfi *et al.*, 2020), adapter ces modèles à des domaines de spécialité se heurte ainsi à la difficulté de trouver des données qualifiées en volume suffisant (Kadhim, 2019; Chawla *et al.*, 2004). Comme les modèles de langue pré-entraînés sur des corpus de langue tous domaines peinent à analyser un vocabulaire et une langue régis par des contraintes spécifiques, propres à un métier ou à une organisation, et différentes de celles des langues naturelles (Ramponi & Plank, 2020; Ji *et al.*, 2021; Pathak *et al.*, 2020), il est donc impératif de trouver un moyen de préparer des données qualifiées en sollicitant des experts de la façon la plus parcimonieuse et optimale possible. L'outil que nous avons développé repose sur une approche hybride. Cette dernière consistant à analyser conjointement le contenu informationnel véhiculé par le texte, d'un côté, et par la structure organisationnelle du texte lui-même, d'un autre côté, permet de construire progressivement des modèles performants d'accès à l'information en impliquant un effort minimal de l'expert. La structure organisationnelle de la documentation est souvent riche en informations, en particulier dans les domaines industriels où la documentation est fortement normalisée. Cette structure traduit une hiérarchisation et une organisation des informations contenues dans les documents selon une logique métier, et qui doit être représentée dans un format exploitable pour des tâches d'extraction et de recherche d'information. Nous proposons une approche originale s'appuyant sur un cycle d'amélioration continue (lifelong learning) (Field, 2000). Afin d'affiner progressivement les prédictions proposées, ces modèles ont été intégrés dans un outil interactif permettant de mettre en oeuvre une forme de supervision faible par des experts dans un cycle itératif de validation des extractions, où la machine apprend à faire des prédictions de plus en plus précises et couvrantes à mesure que de nouveaux exemples et contre-exemples sont collectés. Les exemples (ou contre exemples) évalués peuvent ainsi servir de nouvelles données d'entraînement pour renforcer la faculté de prédiction du système.

L'approche que nous avons mis en oeuvre exploite différents axes de lecture : l'axe de la structure de la documentation (la façon dont les informations sont organisées et découpées à l'intérieur des documents) et celui de la logique métier (information dépendante du domaine d'utilisation). Nous avons utilisé des modèles d'apprentissage supervisés permettant de construire des représentations liées à la fois à la mise en forme des documents et à leur contenu informationnel. Ainsi, l'interface du système permet à la fois d'accéder aux informations pertinentes dans la documentation, corriger les prédictions du système lorsque celui-ci se trompe, entraîner de nouveaux modèles ou ré-entraîner des modèles existants et enfin, construire des données qualifiée pour l'entraînement de nouveaux modèles.

Références

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CHAWLA N. V., JAPKOWICZ N. & KOTCZ A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, **6**(1), 1–6.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FIELD J. (2000). *Lifelong learning and the new educational order*. ERIC.

JI S., HÖLTTÄ M. & MARTTINEN P. (2021). Does the magic of bert apply to medical code assignment ? a quantitative study. *arXiv preprint arXiv :2103.06511*.

KADHIM A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, **52**(1), 273–292.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

PATHAK A. R., AGARWAL B., PANDEY M. & RAUTARAY S. (2020). Application of deep learning approaches for sentiment analysis. In *Deep Learning-Based Approaches for Sentiment Analysis*, p. 1–31. Springer.

RAMPONI A. & PLANK B. (2020). Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6838–6855.

SALLOUM S. A., KHAN R. & SHAALAN K. (2020). A survey of semantic analysis approaches. In *Joint European-US Workshop on Applications of Invariance in Computer Vision*, p. 61–70 : Springer.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

TORFI A., SHIRVANI R. A., KENESHLOO Y., TAVVAF N. & FOX E. A. (2020). Natural language processing advancements by deep learning : A survey. *arXiv preprint arXiv :2003.01200*.