Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Lucie GIANOLA : lucie.gianola@yahoo.fr

Titre : Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique

Mots-clés : reconnaissance d'entités nommées, genre textuel, analyse criminelle, linguistique de corpus.

Title: *Textual Aspects of Judicial Proceedings Files and Perspectives for its Automatic Processing*

Keywords: *named entity recognition, textual genre, criminal analysis, corpus linguistics.*

Thèse de doctorat en sciences du langage, Agora, Université de Cergy-Pontoise, sous la direction de Julien Longhi (Pr, Université de Cergy-Pontoise). Thèse soutenue le 28/02/2020.

Jury : M. Julien Longhi (Pr, Université de Cergy-Pontoise, directeur), M. Patrick Paroubek (IR HDR, LIMSI, CNRS, rapporteur), Mme Sylvie Monjean-Decaudin (Pr, Université Paris IV, rapporteuse et présidente), M. Laurent Chartier (Colonel de gendarmerie, Gendarmerie nationale, examinateur), Mme Bénédicte Pincemin (CR, IHRIM, examinatrice), M. Olivier Ribaux (Pr, Université de Lausanne, examinateur).

Résumé : *L'analyse criminelle est une discipline d'appui aux enquêtes pratiquée au sein de la Gendarmerie nationale. Elle repose sur l'exploitation des documents compilés dans le dossier de procédure judiciaire (auditions, perquisitions, rapports d'expertise, données téléphoniques et bancaires, etc.) afin de synthétiser les informations collectées et de proposer un regard neuf sur les faits examinés. Si l'analyse criminelle a recours à des logiciels de visualisation de données (p. ex.* Analyst's Notebook

d'IBM) pour la mise en forme des hypothèses formulées, la gestion informatique et textuelle des documents de la procédure est à l'heure actuelle entièrement manuelle. D'autre part, l'analyse criminelle s'appuie entre autres sur la conceptualisation des informations du dossier en entités criminelles pour formaliser son travail. La présentation du contexte de recherche détaille la pratique de l'analyse criminelle ainsi que la constitution du dossier de procédure judiciaire en tant que corpus textuel. Nous proposons des perspectives pour l'adaptation des méthodes de traitement automatique de la langue et d'extraction d'information au cas d'étude, notamment la mise en parallèle des concepts d'entité en analyse criminelle et d'entité nommée. Cette comparaison est réalisée sur les plans conceptuels et linguistiques. À l'aide d'un corpus d'exemples constitué d'un dossier de procédure consacré à un homicide, une première approche minimale de détection des entités dans les auditions de témoins basée sur des grammaires locales est présentée. Enfin, le genre textuel étant un paramètre à prendre en compte lors de l'application de traitements automatiques à du texte, nous construisons une structuration du genre textuel « légal » en discours, genres et sous-genres par le biais d'une étude textométrique visant à caractériser différents types de textes (dont les auditions de témoins) produits par le domaine de la justice.

L'objectif de ce travail de thèse est de proposer une compréhension approfondie des concepts des trois domaines concernés (analyse criminelle, extraction d'information et linguistique textuelle) afin de poser les bases méthodologiques et épistémologiques de l'application des méthodes automatiques au cas de l'analyse criminelle.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02522680

Fadila TALEB : talebfadila@gmail.com

Titre : L'argumentation judiciaire à travers le prisme des scénarios modaux

Mots-clés : scénario modal, modalité, zone modale, sémantique des modalités, discours judiciaire, argumentation, rhétorique, genre textuel, textométrie, linguistique de corpus, droit des transports.

Title: *Judicial Argumentation through the Prism of Modal Scenarios. Application for Assistance in the Interpretation of Court Decisions*

Keywords: *modal scenario, modality, modal zone, semantics of modalities, legal discourse, argumentation, rhetoric, kind of text, textometry, corpus linguistics, transportation law.*

Thèse de doctorat en sciences du langage, linguistique, DYnamique du Langage In Situ (DYLIS), département sciences du langage et de la communication, UFR lettres et sciences humaines, Université de Rouen Normandie, sous la direction de Laurent Gosselin (Pr, Université de Rouen Normandie) et Maryvonne Holzem (MC HDR émérite, Université de Rouen Normandie). Thèse soutenue le 8/11/2019.

Jury : M. Laurent Gosselin (Pr, Université de Rouen Normandie, codirecteur), Mme Maryvonne Holzem (MC HDR émérite, Université de Rouen Normandie, codirectrice), M. Laurent Gautier (Pr, Université de Bourgogne, rapporteur), M. Doninique Legallois (Pr, Université Sorbonne Nouvelle-Paris 3, rapporteur), M. Alain Rabatel (Pr, Université Claude Bernard – Lyon 1, président).

Résumé : *Le travail de recherche présenté dans cette thèse s'inscrit dans le cadre général des travaux sur les humanités numériques qui cherchent, entre autres, à contribuer à l'amélioration des interactions homme-machine. L'objectif de l'étude est double. Dans un premier temps, il s'agit d'étudier un corpus de décisions de justice contenues dans la base de données de l'Institut du Droit International des Transports (IDIT) afin de déterminer les contraintes linguistiques du genre judiciaire. Dans un second temps, il est question de proposer des parcours interprétatifs pouvant aider les utilisateurs dans leur accès à l'information juridique recherchée. La problématique de l'aide à l'interprétation est appréhendée à travers l'étude des modalités et des scénarios modaux.*

Le parti pris de cette recherche est de considérer la pluridisciplinarité comme un atout théorique et méthodologique qui contribue à mieux éclairer un objet d'étude. De ce fait, plusieurs approches (sémantique des modalités, sémantique textuelle, argumentation rhétorique, textométrie) sont convoquées et articulées pour œuvrer ensemble vers les objectifs fixés. L'analyse du corpus a été menée à deux niveaux et selon deux approches.

Dans la première partie, l'analyse empirique proposée est quantitative et contrastive. Elle est menée au niveau microtextuel et mésotextuel dans la mesure où elle se focalise sur l'étude du lexique. Aidée de l'outil TXM, cette première investigation a permis une caractérisation linguistique globale du corpus et un premier aperçu de son profil modal grâce notamment à l'introduction de la notion de zone modale. Elle a également mis en exergue des expressions modales, constructions concessives, routines discursives, etc. qui focalisent sur des moments clés dans le déroulement argumentatif et peuvent donc servir dans le cadre de l'aide à l'interprétation.

Dans la seconde partie, l'étude empirique porte sur des analyses modales menées sur des textes complets. Elle est donc abordée dans une approche qualitative et au niveau macrotextuel. Cette analyse aboutit à la formulation d'un modèle de scénario modal minutieusement décrit pour trois sous-genres judiciaires : jugement du tribunal de commerce, arrêt de la cour d'appel et arrêt de la Cour de cassation. Pour chacun des sous-genres, le scénario modal a été décomposé en plusieurs niveaux : scénario modal apparent et scénario modal sous-jacent (selon les modalités qui l'ont construit : modalités de premier plan et modalités d'arrière-plan), et selon qu'il caractérise un texte complet (scénario modal global) ou une zone spécifique de ce texte (sous-scénario modal). Par ailleurs, la présentation schématique (semblable à un algorithme) proposée

pour les scénarios modaux a mis en évidence le rôle que représenterait chaque zone modale dans la perspective d'une aide à l'interprétation.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02995083

Tian TIAN : tian.tian@live.cn

Titre : Adaptation au domaine et combinaison de modèles pour l'annotation de textes multisources et multidomaines

Mots-clés : adaptation au domaine, reconnaissance des entités nommées, apprentissage automatique, champs aléatoires conditionnels, réseaux de neurones.

Title: *Domain Adaptation and Model Combination for the Annotation of Multi-source, Multi-domain Texts*

Keywords: *domain adaptation, named entity recognition, machine learning, conditional random fields, neural networks.*

Thèse de doctorat en sciences du langage, LaTTiCe, UMR 8094, Université Sorbonne Nouvelle – Paris 3, sous la direction de Isabelle Tellier (Pr, Université Sorbonne Nouvelle – Paris 3), Thierry Poibeau (DR, CNRS, LaTTiCe, UMR 8094) et Marco Dinarelli (CR, CNRS, Laboratoire d'Informatique de Genoble, UMR 5217). Thèse soutenue le 16/10/2019.

Jury : Mme Isabelle Tellier (Pr, Université Sorbonne Nouvelle – Paris 3, codirectrice), M. Thierry Poibeau (DR, CNRS, LaTTiCe, UMR 8094, codirecteur), M. Marco Dinarelli (CR, CNRS, Laboratoire d'Informatique de Genoble, UMR 5217, codirecteur), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, rapporteuse), Mme Anne-Laure Ligozat (MC HDR, ENSIIE, LIMSI, rapporteuse),, Mme Sophie Prévost (DR, CNRS, LaTTiCe, UMR 8094, présidente), M. Patrick Marty (IR, Fnac, examinateur).

Résumé : *Aujourd'hui, de nombreux services en ligne proposent aux utilisateurs de commenter, éditer et partager leurs points de vue sur différents sujets de discussion. Ce type de contenu, ou « contenu généré par utilisateur (*user generated content*, UGC) », est maintenant devenu la ressource principale pour les analyses d'opinions sur Internet. Les sujets de ces opinions varient du personnage politique, au produit sur un marché quelconque, au climat, aux sites touristiques et à la vie privée personnelle. L'analyse de cette masse de données permet de suivre l'évolution des opinions au fil du temps, ce qui pourrait être utile pour les changements de stratégies ou le choix de décision, et permettre l'évaluation de ces changements et des prises de décisions. Néanmoins, à cause des abréviations, du bruit, des fautes d'orthographe et de toute autre sorte de problèmes, les outils de classifications des textes et de traitements automatiques des langues ont des performances plus faibles que sur les textes bien formés.*

Cette thèse a pour objet la reconnaissance d'entités nommées sur les contenus générés par les utilisateurs sur Internet, les données cibles viennent essentiellement de forums spécialisés, Facebook et Twitter. Nous avons établi un corpus d'évaluation avec des textes multisources et multidomaines : nous avons distingué les textes de forums, de Facebook et des tweets *comme différentes sources et nous avons traité les domaines de discussions sur des produits qui varient entre fast-food, automobile, musique en* streaming *et jouets pour les enfants. Ensuite, nous avons développé un modèle de champs aléatoires conditionnels, entraîné sur un corpus annoté disponible, provenant des contenus générés par les utilisateurs (uniquement extrait de Twitter).*

Dans le but d'améliorer les résultats de la reconnaissance d'entités nommées, nous avons d'abord développé un étiqueteur morphosyntaxique pour les contenus générés par les utilisateurs. Les étiqueteurs morphosyntaxiques appris sur les textes bien formés ne fonctionnent pas aussi bien sur les données générées par les utilisateurs. Nous avons donc testé deux méthodes d'adaptation au domaine en mélangeant les textes bien formés et les textes générés par les utilisateurs dans les données d'apprentissage pour améliorer la performance de l'étiqueteur morphosyntaxique. Ensuite, nous avons utilisé les étiquettes prédites par cet étiqueteur morphosyntaxique comme un attribut du modèle des champs aléatoires conditionnels pour le modèle d'extraction d'entités nommées. Enfin, pour transformer les contenus générés par les utilisateurs en textes bien formés, nous avons développé un modèle de normalisation lexicale basé sur des réseaux de neurones pour détecter les mots non standards et proposer une forme correcte pour les remplacer.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02473489/