Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

INRIA, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
sylvain.pogodalla@inria.fr

Marion BARANES : marion.baranes@gmail.com

Titre : Normalisation orthographique de corpus bruités

Mots-clés : Normalisation, correction orthographique, mots inconnus, altérations, données produites par l'utilisateur.

Title: *Spelling Normalisation of Noisy Text*

Keywords: *Normalization, spell-checking, unknown-word, spelling mistake, user-generated content, natural language processing.*

Thèse de doctorat en Linguistique Théorique, Descriptive et Automatique, UFR Linguistique, Université Paris Diderot – Paris 7, sous la direction de Laurence Danlos (Pr, Université Paris Diderot – Paris 7) et Benoît Sagot (CR, INRIA Paris – Rocquencourt). Thèse soutenue le 23/10/2015.

Jury : Mme Delphine Bernhard (MC, Université de Strasbourg, examinatrice), Mme Laurence Danlos (Pr, Université Paris Diderot – Paris 7, codirectrice), M. Cédrick Fairon (Pr, Université catholique de Louvain, Belgique, rapporteur et président), M. Philippe Langlais (Pr, Université de Montréal, Canada, rapporteur), M. Benoît Sagot (CR, INRIA Paris – Rocquencourt, codirecteur).

Résumé : *L'information contenue dans les messages publiés par les internautes (forums, réseaux sociaux, sites d'avis, etc.) comporte un intérêt stratégique pour de nombreuses entreprises. Néanmoins, peu d'outils ont été conçus pour faciliter l'analyse de ces messages, dont l'orthographe, la typographie et la syntaxe sont souvent bruitées.*

Cette thèse industrielle a été réalisée au sein de l'entreprise viavoo afin d'améliorer les résultats d'un outil d'extraction d'information qui fait abstraction de la variabilité flexionnelle. Nous avons ainsi développé une chaîne de traitements pour la normalisation orthographique de textes bruités. Son objectif est donc de transformer ces textes pour faire en sorte que tous les mots qui les composent obtiennent une orthographe standard, à la flexion près.

L'approche présentée ici consiste tout d'abord à déterminer automatiquement, parmi les tokens du corpus traité qui sont inconnus d'un lexique de référence, ceux qui résultent d'altérations et qu'il conviendrait donc de normaliser, par opposition aux autres (néologismes, emprunts, etc.). Des candidats de normalisation sont alors proposés pour ces tokens à l'aide de règles pondérées obtenues par des techniques d'apprentissage par analogie. Nous identifions ensuite des tokens connus du lexique de référence mais qui résultent néanmoins d'une altération (fautes grammaticales), et proposons des candidats de normalisation pour ces tokens. Enfin, des modèles de langue permettent de prendre en compte le contexte dans lequel apparaissent les différents types d'altérations pour lesquels des candidats de normalisation ont été proposés afin de choisir les plus probables.

Différentes expériences et évaluations sont réalisées sur le français à chaque étape et sur la chaîne complète. Une attention particulière a été portée au caractère faiblement dépendant de la langue des modules développés, ce qui permet d'envisager son adaptation à d'autres langues européennes (ex. : anglais, allemand ou encore espagnol).

URL où le mémoire pourra être téléchargé :
https://hal.inria.fr/tel-01226159

Chloé BRAUD : chloe.braud@gmail.com

Titre : Identification automatique des relations discursives implicites à partir de corpus annotés et de données brutes

Mots-clés : Relations discursives implicites, analyse discursive, apprentissage automatique, apprentissage avec données brutes.

Title: *Automatic Identification of Implicit Discourse Relations from Annotated Corpora and Raw Data*

Keywords: *Implicit discourse relations, discourse analysis, machine learning, learning with raw data.*

Thèse de doctorat en Sciences du Langage, UFR Sciences du Langage, Université Paris Diderot – Paris 7, sous la direction de Laurence Danlos (Pr, Université Paris Diderot – Paris 7) et Pascal Denis (CR, INRIA Lille – Nord-Europe). Thèse soutenue le 18/12/2015.

Jury : Mme Laurence Danlos (Pr, Université Paris Diderot – Paris 7, codirectrice), M. Pascal Denis (CR, INRIA Lille – Nord-Europe, codirecteur), Mme Liesbeth Degand (Pr, Université catholique de Louvain, Belgique, présidente), M. Philippe Muller (MC, Université Toulouse III – Paul Sabatier, IRIT, rapporteur), Mme Caroline Sporleder (Pr, Georg-August-Universität Göttingen, Allemagne, rapporteur).

Résumé : *Le développement de systèmes d'analyse discursive automatique des documents est un enjeu actuel majeur en Traitement Automatique des Langues. La difficulté principale correspond à l'étape d'identification des relations (comme* Explication*,* Contraste*. . .) liant les segments constituant le document. En particulier, l'identification des relations dites implicites, c'est-à-dire non marquées par un connecteur discursif (comme* mais*,* parce que*. . .), est réputée difficile car elle nécessite la prise en compte d'indices variés et correspond à des difficultés particulières dans le cadre d'un système de classification automatique. Dans cette thèse, nous utilisons des données brutes pour améliorer des systèmes d'identification automatique des relations implicites.*

Nous proposons d'abord d'utiliser les connecteurs pour annoter automatiquement de nouvelles données. Nous mettons en place des stratégies issues de l'adaptation de domaine qui nous permettent de gérer les différences en termes distributionnels entre données annotées automatiquement et manuellement : nous rapportons des améliorations pour des systèmes construits sur le corpus français ANNODIS et sur le corpus anglais du Penn Discourse Treebank. Ensuite, nous proposons d'utiliser des représentations de mots acquises à partir de données brutes, éventuellement annotées automatiquement en connecteurs, pour enrichir la représentation des données fondées sur les mots présents dans les segments à lier. Nous rapportons des améliorations sur le corpus anglais du Penn Discourse Treebank et montrons notamment que cette méthode permet de limiter le recours à des ressources riches, disponibles seulement pour peu de langues.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01256884

Iris ESHKOL-TARAVELLA : iris.eshkol@univ-orleans.fr

Titre : La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral

Mots-clés : Annotation, corpus oral, TAL, étiquetage morpho-syntaxique, chunking, reformulation, entités nommées, anonymisation, noms généraux, noms de lieux, lieux subjectifs, subjectivité, recettes de cuisine.

Title: *Specification of Linguistic Annotations According to Corpora: from Newspaper to Spoken Corpora*

Keywords: *Annotation, oral corpus, NLP, POS tagging, chunking, reformulation, named entities, anonymisation, general nouns, toponyms, subjective places, subjectivity, recipes.*

Habilitation à diriger des recherches en Sciences du Langage, UFR Collegium d'ITP Lettres, Langues et Sciences Humaines, Université d'Orléans, sous la direction de Gabriel Bergounioux (Pr, Université d'Orléans, LLL-CNRS). Habilitation soutenue le 16/10/2015.

Jury : M. Gabriel Bergounioux (Pr, Université d'Orléans, LLL-CNRS, directeur), Mme Isabelle Tellier (Pr, Université Paris 3 – Sorbonne Nouvelle, LaTTiCe-CNRS, UMR 8094, présidente), Mme Catherine Schnedecker (Pr, Université de Strasbourg, LiLPa, rapporteur), Zweigenbaum Pierre (DR, CNRS, LIMSI, Orsay, rapporteur), M. Massimo Moneglia (Pr, Università degli Studi di Firenze, LABLITA, Florence, Italie, rapporteur), M. Denis Maurel (Pr, Université François Rabelais, Tours, examinateur).

Résumé : *Confronté à Internet, le Traitement Automatique des Langues (TAL) a dû relever le défi que posait l'analyse de textes dialogiques écrits (blog, forum, chat, réseaux sociaux, etc.) et oraux. Les recherches présentées ont, dans un premier temps, porté sur le développement de systèmes à même de repérer et d'analyser l'information à partir d'une annotation des ressources. L'approche retenue privilégie l'intégration d'indices inhérents à la nature de corpus « hors normes » afin d'améliorer les techniques de traitement automatique.*

La chaîne d'opérations comprend quatre étapes :

(i) L'observation et l'analyse manuelle des données afin de recenser les variations dans les occurrences et d'évaluer l'ampleur des phénomènes à annoter, leur classification et l'identification de leurs marqueurs formels.

(ii) La modélisation de l'information à partir d'une typologie sous la forme d'un jeu d'étiquettes ajusté à la nature du corpus.

(iii) La définition de la technologie congrue (généralement, l'arbitrage entre le développement d'un nouvel outil et l'adaptation d'un outil existant).

(iv) L'implémentation du schéma d'annotation défini afin de procéder à une analyse quantitative et qualitative des résultats.

L'annotation effectuée concerne les domaines de la syntaxe (étiquetage morpho-syntaxique et chunking), sémantique et/ou pragmatique (entités nommées, indices d'identification de la personne, reformulations, etc.). L'application concerne aussi bien des entretiens transcrits que des titres de cartes géographiques, des recettes d'omelette que des articles du Monde*. Les méthodes utilisées varient en fonction du corpus et de la tâche traitée. L'annotation syntaxique et le repérage des segments reformulés sont fondés sur la technique d'apprentissage automatique avec les CRFs ; le repérage des entités nommées et des indices d'identification de la personne dans les transcriptions de l'oral utilise les méthodes symboliques ; la détection automatique des tours de parole contenant la reformulation emploie les méthodes heuristiques.*

Le travail sur le français parlé et son annotation a conduit à la modélisation des caractéristiques propres à l'oral : disfluences, marqueurs discursifs, présentateurs, segmentation, commentaires personnels, etc. Un autre phénomène caractéristique de l'oral, la reformulation, a fait l'objet d'une étude particulière.

Le travail sur l'annotation du corpus oral, du corpus Web ou du corpus médiatique a permis de reconsidérer la notion de subjectivité qui constitue l'une des difficultés récurrentes du traitement automatique. L'étude de la subjectivité et son expression dans le discours a été poursuivie dans plusieurs des recherches menées : la subjectivité à partir des informations personnelles livrées par le locuteur, la subjectivité dans la perception et l'appropriation des lieux, la subjectivité dans les recettes de cuisine et enfin la subjectivité exprimée à travers les noms généraux.

URL où le mémoire pourra être téléchargé :
https://hal.archives-ouvertes.fr/tel-01250650

Jérôme KIRMAN : jkirman@labri.fr

Titre : Mise au point d'un formalisme de haut niveau pour le traitement automatique des langues

Mots-clés : Linguistique informatique, logique, lambda-calcul, grammaires catégorielles abstraites, syntaxe modèle-théorique.

Title: *A High-Level Formalism for Natural Language Processing*

Keywords: *Computational linguistics, logic, lambda-calculus, abstract categorial grammars, model-theoretic syntax.*

Thèse de doctorat en Informatique, LaBRI – UMR 5800, Université de Bordeaux, sous la direction de Bruno Courcelle (Pr émérite, Université de Bordeaux, LaBRI), Sylvain Salvati (CR, INRIA Bordeaux – Sud Ouest) et Lionel Clément (MC, Université de Bordeaux, LaBRI). Thèse soutenue le 04/12/2015.

Jury : M. Géraud Sénizergues (Pr, Université de Bordeaux, LaBRI, président), M. Bruno Courcelle (Pr émérite, Université de Bordeaux, LaBRI, codirecteur), M. Sylvain Salvati (CR, INRIA Bordeaux – Sud Ouest, codirecteur), M. Lionel Clément (MC, Université de Bordeaux, LaBRI, codirecteur), M. Guy Perrier (Pr émérite, Université de Lorraine, LORIA, UMR 7503, Nancy, rapporteur), M. Denys Duchier (Pr, Université d'Orléans, LIFO, rapporteur), M. Éric Villemonte de la Clergerie (CR, INRIA Paris – Rocquencourt, examinateur), M. Sylvain Schmitz (MC, ENS Cachan, examinateur).

Résumé : *La linguistique informatique a pour objet de construire un modèle formel des connaissances linguistiques, et d'en tirer des algorithmes permettant le traitement automatique des langues. Pour ce faire, elle s'appuie fréquemment sur des grammaires dites génératives, construisant des phrases valides par l'application suc-*

cessive de règles de réécriture. Une approche alternative, basée sur la théorie des modèles, vise à décrire la grammaticalité comme une conjonction de contraintes de bonne formation, en s'appuyant sur des liens profonds entre logique et automates pour produire des analyseurs efficaces. Notre travail se situe dans ce dernier cadre.

En s'appuyant sur plusieurs résultats existant en informatique théorique, nous proposons un outil de modélisation linguistique expressif, conçu pour faciliter l'ingénierie grammaticale. Il considère dans un premier temps la structure abstraite des énoncés, et fournit un langage logique s'appuyant sur les propriétés lexicales des mots pour caractériser avec concision l'ensemble des phrases grammaticalement correctes. Puis, dans un second temps, le lien entre ces structures abstraites et leurs représentations concrètes (en syntaxe et en sémantique) est établi par le biais de règles de linéarisation qui exploitent la logique et le lambda-calcul.

Par suite, afin de valider cette approche, nous proposons un ensemble de modélisations portant sur des phénomènes linguistiques divers, avec un intérêt particulier pour le traitement des langages présentant des phénomènes d'ordre libre (c'est-à-dire qui autorisent la permutation de certains mots ou groupes de mots dans une phrase sans affecter sa signification), ainsi que pour leur complexité algorithmique.

URL où le mémoire pourra être téléchargé :
https://hal.archives-ouvertes.fr/tel-01251668

Nada MIMOUNI : nada.mimouni@gmail.com, nada.mimouni@lipn.univ-paris13.fr

Titre : Interrogation d'un réseau sémantique de documents : l'intertextualité dans l'accès à l'information juridique

Mots-clés : Réseau de documents, intertextualité, recherche d'information, analyse formelle et relationnelle de concepts, web sémantique, ontologies.

Title: *Querying a Semantic Network of Documents: Intertextuality in Legal Information Access*

Keywords: *Document network, intertextuality, information retrieval, formal and relational concept analysis, semantic web, ontologies.*

Thèse de doctorat en Informatique, Groupe Mathématique, Informatique, Signal, École doctorale Galilée, Laboratoire Informatique de Paris Nord (LIPN) – CNRS UMR 7030, Université Paris Nord – Paris 13, Villetaneuse, sous la direction de Adeline Nazarenko (Pr, Université Paris Nord – Paris 13, LIPN) et Sylvie Salotti (MC, Université Paris Nord – Paris 13, LIPN). Thèse soutenue le 27/01/2015.

Jury : Mme Adeline Nazarenko (Pr, Université Paris Nord – Paris 13, LIPN, codirectrice), Mme Sylvie Salotti (MC, Université Paris Nord – Paris 13, LIPN, codirectrice), Mme Sylvie Calabretto (Pr, INSA de Lyon, rapporteur), M. Ollivier Haemmerlé (Pr, Université Toulouse – Jean Jaurès, rapporteur), Mme Danièle Bourcier (DR, CNRS,

CERSA, Paris, examinatrice), M. Aldo Gangemi (Pr, Université Paris Nord – Paris 13, LIPN, examinateur), M. Amedeo Napoli (DR, CNRS, LORIA, UMR 7503, Nancy, président), Mme Chantal Reynaud (Pr, Université Paris Sud – Paris 11, Orsay, examinatrice).

Résumé : *La recherche d'information considère généralement les documents comme des unités indépendantes. Le modèle traditionnel de recherche d'information ne prend pas en compte la richesse du réseau des relations sémantiques qui peuvent structurer les collections documentaires.*

Dans le domaine juridique, cette limitation est critique du fait de l'abondance et de la diversité des relations qui lient entre elles les sources de loi. En effet, l'intertextualité est au cœur de toute activité juridique (production de textes de lois, modification, etc.). Elle est reconnue comme un facteur majeur de complexité qui doit être pris en compte dans le processus d'accès à l'information. Ceci a été confirmé par l'analyse des besoins des professionnels de droit : ceux-ci cherchent à formuler des requêtes complexes qui portent aussi bien sur le contenu des documents que sur les relations intertextuelles qu'ils entretiennent.

Nous proposons un nouveau modèle de recherche d'information où les collections documentaires sont modélisées comme des graphes de documents attribués et l'interrogation comme un appariement de graphes. Ce modèle logique est bien adapté aux professionnels du droit qui ont besoin de retrouver tous les documents relatifs à un cas et pas seulement les plus pertinents. Le graphe d'une collection documentaire rend compte du contenu sémantique des documents, de leurs types et des relations intertextuelles qu'ils entretiennent.

Nous avons proposé deux approches pour implémenter ce modèle. Elles permettent de répondre à des requêtes relationnelles et de retourner en réponse des graphes de documents. La première approche est structurée. Elle utilise l'analyse formelle et l'analyse relationnelle de concepts pour construire une structure conceptuelle (une famille relationnelle de treillis de concepts documentaires) au-dessus du graphe de la collection documentaire. Nous avons défini des méthodes de recherche de documents par accès direct ou par navigation pour interroger et explorer le modèle relationnel construit puis retourner des documents ou graphes de documents pertinents. La deuxième approche interroge directement le graphe documentaire. Elle présente une solution plus opérationnelle qui repose sur les technologies du web sémantique et un modèle à base d'ontologie. Les collections sont modélisées comme des ensembles de documents liés. L'interrogation est faite à l'aide de SPARQL.

En recherche d'information, structurer une collection de documents sous forme de treillis pré-calcule les réponses à toutes les requêtes satisfiables sur cette collection. De plus, naviguer dans la structure fournit des réponses approximatives en généralisant ou spécifiant la requête d'origine. Dans l'approche directe, les réponses sont calculées à la volée lorsque la requête est envoyée au système. Cette approche est plus flexible (lorsque le modèle de la collection ou la collection elle-même évolue), n'est pas limitée par la taille de la collection et permet d'exprimer des requêtes plus

riches. À l'inverse, l'approche structurée obtient de bons résultats sur des petites collections ou des perspectives locales restreintes à des sous-collections de collections plus larges et permet la navigation et la visualisation, mais ne peut pas répondre à des requêtes complexes.

Le choix d'utiliser une approche ou l'autre dépend étroitement des besoins de l'application (les interfaces utilisateur, le nombre de documents, la granularité de description de document ou l'évolution de la collection). En perspective, nous explorons la piste d'implémenter les stratégies de navigation pertinentes sous forme de séquences de requêtes SPARQL et de concevoir des interfaces pour guider les utilisateurs dans la création de requêtes relationnelles complexes et analyser les résultats retournés.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01230641

Jorge Mauricio MOLINA MEJIA : jorge.mauricio.molina@gmail.com

Titre : ELiTe-[FLE]^2 : un environnement d'ALAO fondé sur la linguistique textuelle, pour la formation linguistique des futurs enseignants de FLE en Colombie

Mots-clés : Formation des enseignants de FLE, linguistique textuelle, ALAO, TAL, corpus.

Title: *ELiTe-[FLE]^2: a CALL Environment Based on Text Linguistics, Aimed at Helping Future FFL Teachers in Colombia*

Keywords: *FFL teacher training, text linguistics, CALL, NLP, corpus.*

Thèse de doctorat en Informatique et Sciences du Langage, Laboratoire de linguistique et didactique des langues étrangères et maternelles — LIDILEM —, UFR Langage, lettres et arts du spectacle, information et communication — LLASIC —, Université Grenoble Alpes, sous la direction de Georges Antoniadis (Pr, Université Stendhal – Grenoble 3). Thèse soutenue le 06/11/2015.

Jury : M. Georges Antoniadis (Pr, Université Stendhal – Grenoble 3, directeur), Mme Lita Sander Lundquist (Pr émérite, Copenhagen Business School, Danemark, rapporteur), M. Jean-Pierre Cuq (Pr, Université de Nice, rapporteur), M. Jean-Marc Colletta (Pr, Université Stendhal – Grenoble 3, président), M. Eric Wehrli (Pr, Université de Genève, Suisse, examinateur).

Résumé : *Nous présentons, dans ce manuscrit, un dispositif informatique d'aide à la formation des futurs enseignants de FLE en Colombie. Il prend ses sources dans la linguistique textuelle et cherche à améliorer le niveau linguistique des étudiants universitaires actuellement en formation. Pour ce faire, le dispositif est fondé sur un corpus textuel spécifiquement annoté et étiqueté grâce aux outils de traitement automatique de langues (TAL) et à des annotations manuelles en format XML. Ceci permet de développer des activités à visée formative, en tenant compte des besoins exprimés par les*

publics cibles (enseignants-formateurs et leurs étudiants en formation). Comme nous l'exposons tout au long de cette thèse, l'élaboration d'un système comme le nôtre est le produit de la mise en œuvre de connaissances et de compétences issues de plusieurs disciplines et domaines : didactique des langues, ingénierie pédagogique, linguistique générale, linguistique textuelle, linguistique de corpus, TAL et ALAO. Il se veut, principalement, un dispositif pédagogique pour la formation des étudiants en FLE dans le contexte de l'éducation supérieure en Colombie, et un outil pensé en fonction des besoins et des objectifs de cet apprentissage. L'originalité de notre système repose sur le type de public choisi, le modèle didactique de formation mis en œuvre et la spécificité du corpus utilisé. À notre connaissance, il s'agit d'un des premiers systèmes d'ALAO fondé sur la linguistique textuelle s'adressant à la formation des futurs enseignants de FLE dans un contexte exolingue.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01228439

Ludovic MONCLA : moncla.ludovic@gmail.com

Titre : Reconstruction automatique d'itinéraire à partir de textes descriptifs

Mots-clés : Extraction d'information, reconstruction automatique d'itinéraire, traitement automatique du language naturel.

Title: *Automatic Reconstruction of Itineraries from Descriptive Texts*

Keywords: *Information extraction, automatic itinerary reconstruction, natural language processing.*

Thèse de doctorat en Informatique, Département Informatique, UFR Sciences et Techniques, Université de Pau et des Pays de l'Adour, sous la direction de Mauro Gaio (Pr, Université de Pau et des Pays de l'Adour, LIUPPA), Javier Nogueras-Iso (Pr, Universidad de Zaragoza, DIIS, Espagne), Sébastien Mustière (Ingénieur des Travaux Géographiques et Cartographiques de l'Etat, HDR, IGN, Conception Objet et Généralisation de l'Information Topographique, COGIT, Saint-Mandé). Thèse soutenue le 03/12/2015.

Jury : M. Mauro Gaio (Pr, Université de Pau et des Pays de l'Adour, LIUPPA, codirecteur), M. Javier Nogueras-Iso (Pr, Universidad de Zaragoza, DIIS, Espagne, codirecteur), M. Sébastien Mustière (Ingénieur des Travaux Géographiques et Cartographiques de l'Etat, HDR, IGN, Conception Objet et Généralisation de l'Information Topographique, COGIT, Saint-Mandé, codirecteur), M. Christophe Claramunt (Pr, Institut de Recherche de l'Ecole Navale, Brest, rapporteur), M. Denis Maurel (Pr, Université François Rabelais, Tours, rapporteur), M. Ross Purves (Pr, University of Zurich, Suisse, rapporteur), Mme Adeline Nazarenko (Pr, Université Paris Nord – Paris 13, LIPN, présidente), M. Philippe Muller (MC, Université Toulouse III – Paul Sabatier, IRIT, examinateur).

Résumé : *Cette thèse s'inscrit dans le cadre du projet PERDIDO dont les objectifs sont l'extraction et la reconstruction d'itinéraires à partir de documents textuels. Ces travaux ont été réalisés en collaboration entre le laboratoire LIUPPA de l'université de Pau et des Pays de l'Adour (France), l'équipe Systèmes d'Information Avancés (IAAA) de Universidad de Zaragoza (Espagne) et le laboratoire COGIT de l'IGN (France). Les objectifs de cette thèse sont de concevoir un système automatique permettant d'extraire, dans des récits de voyages ou des descriptions d'itinéraires, des déplacements, puis de les représenter sur une carte.*

Nous proposons une approche automatique pour la représentation d'un itinéraire décrit en langage naturel. Notre approche est composée de deux tâches principales. La première tâche a pour rôle d'identifier et d'extraire les informations qui décrivent l'itinéraire dans le texte, comme par exemple les entités nommées de lieux et les expressions de déplacement ou de perception. La seconde tâche a pour objectif la reconstruction de l'itinéraire. Notre proposition combine l'utilisation d'informations extraites grâce au traitement automatique du langage ainsi que des données extraites de ressources géographiques externes (comme des gazetiers).

L'étape d'annotation d'informations spatiales est réalisée par une approche qui combine l'étiquetage morpho-syntaxique et des patrons lexico-syntaxiques (cascade de transducteurs) afin d'annoter des entités nommées spatiales et des expressions de déplacement ou de perception. Une première contribution au sein de la première tâche est la désambiguïsation des toponymes, qui est un problème encore mal résolu en reconnaissance d'entités nommées et essentiel en recherche d'information géographique. Nous proposons un algorithme non-supervisé de géo-référencement basé sur une technique de clustering capable de proposer une solution pour désambiguïser les toponymes trouvés dans les ressources géographiques externes, et dans le même temps proposer une estimation de la localisation des toponymes non référencés.

Nous proposons un modèle de graphe générique pour la reconstruction automatique d'itinéraire, où chaque nœud représente un lieu et chaque segment représente un chemin reliant deux lieux.

L'originalité de notre modèle est qu'en plus de tenir compte des éléments habituels (chemins et points de passage), il permet de représenter les autres éléments impliqués dans la description d'un itinéraire, comme par exemple les points de repère visuels. Un calcul d'arbre de recouvrement minimal à partir d'un graphe pondéré est utilisé pour obtenir automatiquement un itinéraire sous la forme d'un graphe. Chaque segment du graphe initial est pondéré en utilisant une méthode d'analyse multi-critère combinant des critères qualitatifs et des critères quantitatifs. La valeur des critères est déterminée à partir d'informations extraites du texte et d'informations provenant de ressources géographiques externes. Par exemple, nous combinons les informations issues du traitement automatique de la langue comme les relations spatiales décrivant une orientation (ex : se diriger vers le sud) avec les coordonnées géographiques des lieux trouvés dans les ressources pour déterminer la valeur du critère « relation spatiale».

De plus, à partir de la définition du concept d'itinéraire et des informations utilisées dans la langue pour décrire un itinéraire, nous avons modélisé un langage d'annotation d'information multi-couche. Ce language s'appuie sur une couche générique basée sur les recommendations du consortium TEI (Text Encoding and Interchange) et peut être adapté en plusieurs couches spécifiques ajoutant de la sémantique aux éléments et aux relations annotées.

Enfin, nous avons implémenté et évalué les différentes étapes de notre approche sur un corpus multilingue de descriptions de randonnées (français, espagnol et italien).

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01249999

Nikola TULECHKI : nikola.tulechki@univ-tlse2.fr

Titre : Traitement automatique de rapports d'incidents et accidents : application à la gestion du risque dans l'aviation civile

Mots-clés : Traitement automatique des langues, retour d'expérience, aviation civile, similarité textuelle, gestion du risque.

Title: *Natural Language Processing of Incident and Accident Reports: Application to Risk Management in Civil Aviation*

Keywords: *Natural language processing, incident reporting, civil aviation, textual similarity, safety management.*

Thèse de doctorat en Sciences du Langage, CLLE-ERSS, Université de Toulouse, sous la direction de Ludovic Tanguy (MC HDR, Université de Toulouse, CLLE-ERSS, UMR 5263). Thèse soutenue le 30/09/2015.

Jury : M. Ludovic Tanguy (MC HDR, Université de Toulouse, CLLE-ERSS, UMR 5263, directeur), M. Yannick Toussaint (CR HDR, INRIA Nancy – Grand-Est, LORIA, UMR 7503, rapporteur), M. Patrice Bellot (Pr, Aix Marseille Université, LSIS, rapporteur), Mme Cécile Fabre (Pr, Université de Toulouse II – Jean Jaurès, examinatrice), M. Éric Hermann (Directeur, CFH/Safety Data Analysis Services, examinateur).

Résumé : *Cette thèse décrit les applications du traitement automatique des langues (TAL) à la gestion des risques industriels. Elle se concentre sur le domaine de l'aviation civile, où le retour d'expérience (REX) génère de grandes quantités de données, sous la forme de rapports d'accidents et d'incidents.*

Nous commençons par faire un panorama des différents types de données générées dans ce secteur d'activité. Nous analysons les documents, comment ils sont produits, collectés, stockés et organisés, ainsi que leurs utilisations. Nous montrons que le paradigme actuel de stockage et d'organisation est mal adapté à l'utilisation réelle de ces

documents et identifions des domaines problématiques où les technologies du langage constituent une partie de la solution.

Répondant précisément aux besoins d'experts en sécurité, deux solutions initiales sont implémentées : la catégorisation automatique de documents afin d'aider le codage des rapports dans des taxonomies préexistantes et un outil pour l'exploration de collections de rapports, basé sur la similarité textuelle.

En nous basant sur des observations de l'usage de ces outils et sur les retours de leurs utilisateurs, nous proposons différentes méthodes d'analyse des textes issus du REX et discutons des manières dont le TAL peut être appliqué dans le cadre de la gestion de la sécurité dans un secteur à haut risque. En déployant et évaluant certaines solutions, nous montrons que même des aspects subtils liés à la variation et à la multidimensionnalité du langage peuvent être traités en pratique afin de gérer la surabondance de données REX textuelles de manière ascendante.

URL où le mémoire pourra être téléchargé :
https://hal.archives-ouvertes.fr/tel-01230079