Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT)

Frédéric Landragin

Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT)

Frédéric Landragin

Laboratoire Lattice, CNRS, Ecole normale supérieure, Université Paris 3, Université Sorbonne Paris Cité, PSL Research University, Paris/Montrouge

http://www.lattice.cnrs.fr/

DRAFT AUTEUR

Nom du projet :

DEMOCRAT : DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique.

Autres partenaires du projet :

Laboratoire LiLPa, « Linguistique, Langues, Parole », EA 1339, Strasbourg.

Laboratoire ICAR, « Interactions, Corpus, Apprentissages, Représentations », UMR 5191, CNRS/ENS Lyon/Lyon 2.

Thématique générale de l'équipe :

Le Lattice est un laboratoire dont les recherches concernent essentiellement la linguistique et le traitement automatique des langues. Les tutelles du laboratoire sont le CNRS, l'Ecole normale supérieure et l'Université Paris 3 Sorbonne Nouvelle, ce qui met le laboratoire au contact direct de nombreuses équipes couvrant tout le spectre de la recherche en littérature et sciences sociales. Le laboratoire est en outre membre de deux labex : le labex Transfers d'une part (labex qui regroupe l'ensemble des laboratoires de recherche en lettres et sciences humaines de l'ENS Ulm) et le labex EFL d'autre part (labex « *Empirical Foundations of Linguistics* », qui regroupe la plupart des laboratoires de sciences du langage de la COMUE Sorbonne Paris Cité). Du fait de cet environnement pluridisciplinaire très riche, le laboratoire est sollicité depuis plusieurs années par des collègues d'autres disciplines ayant des besoins particuliers pour structurer, analyser ou valoriser de grandes masses de données textuelles dans des domaines très variés.

Certaines demandes peuvent être traitées à partir d'outils standard. Par exemple, un besoin très répandu concerne la mise en ligne de documents uniquement disponibles sous forme papier. Au-delà de l'aspect numérisation, la valorisation de ces documents passe généralement par une ré-analyse du contenu, c'est-à-dire l'extraction et la normalisation du vocabulaire technique, sa structuration, la mise en place d'index structurés, etc. Les outils d'extraction de termes, de structuration des connaissances et de mise en place d'hypertextes sont lors sollicités. Une adaptation au contexte est toutefois systématiquement nécessaire, ainsi que la collaboration avec des experts du

domaine visé. Des expériences récentes ont par exemple eu lieu avec des textes d'archéologie (collaboration autour du projet PEPS EITAB porté par le laboratoire AOROC de l'ENS) : le repérage de termes du domaine de l'archéologie, ainsi que la reconnaissance des entités nommées sont très utiles [Mélanie-Becquet, 2015]. A partir du PDF reflétant le texte original, il est possible de produire des bases de données indexant les découvertes archéologiques par commune, par type ou par période, ce qui permet de concevoir des requêtes d'une richesse incomparable par apport à un simple support papier. Un expert du domaine est bien évidemment nécessaire pour valider les résultats des outils automatiques, structurer les données (repérer les synonymes, les hyperonymes, structurer les connaissances) mais le traitement d'un ouvrage complet (c'est-à-dire le passage du papier au support informatique) est ainsi possible en quelques jours à peine.

Un projet peut-être plus original est né d'une collaboration avec nos collègues londoniens du laboratoire d'Humanités numériques de University College London (UCL). Cette université dispose d'une collection de manuscrits de Jeremy Bentham dont une partie est toujours inédite. Ces manuscrits inédits one été transcrits par une équipe de volontaires, ce qui a permis d'obtenir un ensemble de 30 000 documents (fichiers informatiques au format XML) dont le contenu était largement inconnu ou, tout au moins, n'a encore fait l'objet d'aucune étude systématique. La masse de documents à traiter, même si chaque fichier est très bref, rend difficile une approche purement manuelle. La stratégie développée à consister à normaliser les contenus, extraire un certain nombre d'expressions clés, puis, sur cette base, à proposer différentes visualisations correspondant à différents regroupements de documents, chaque regroupement pouvant en outre recevoir une étiquette rendant compte de son contenu. Les experts ont ainsi un accès beaucoup plus facile au corpus, même si les documents restent à analyser plus finement par des experts du philosophe. Il ne s'agit en aucun cas de se substituer au spécialiste du domaine mais de lui donner les clés pour accéder rapidement à l'information qu'il cherche, aux documents les plus proches ou aux principaux thèmes abordés dans un texte u, en l'occurrence dans un grand corpus inédit.

La linguistique est elle-même demandeuse d'automatisation, pour l'annotation de gros corpus en particulier. Le Lattice est impliqué dans plusieurs projets visant à ajouter des informations de nature diverses (morphosyntaxiques, syntaxiques, voire sémantiques) sur des corpus de langues diverses. Citons par exemple le projet DFG-ANR SRCMF (*Syntactic Reference Corpus of Medieval* French) qui visait à produire un corpus d'ancien français annoté au niveau syntaxique afin de faire progresser notre connaissance de l'ancien français et surtout de son évolution au cours des siècles [Prévost, 2015]. Le corpus résultant (corpus de plus de 200 000 mots, interrogeable en ligne) permet aujourd'hui des analyses systématiques et surtout quantifiées qui n'étaient pas possibles jusqu'ici.

Un autre projet, financé par PSL (Paris Sciences et Lettres) cette fois-ci, se met actuellement en place pour l'annotation de textes en hébreu rabbinique, en ancien français et dans plusieurs langues finno-ougriennes. Ce projet appelé Lakmé, et mené en collaboration avec des partenaires de l'Ecole Pratique des Hautes Etudes (EPHE) et de l'Ecole Nationale des Chartes (ENC) demande l'exploration de techniques d'annotation innovantes : l'hébreu comme les langues finno-ougriennes sont des langues « agglutinantes », ce qui signifie que les mots comportent une morphologie très riche et porteuse d'informations essentielles pour l'annotation. Les techniques classiques développées initialement pour l'anglais ne sont pas opérationnelles dans ce cadre et une

approche complètement nouvelle doit être mise au point. Ce projet a, de plus, un but directement pratique : il s'agit de faciliter le travail des experts de ces langues dans leur exploration de grands corpus, afin de mettre au jour des tendances, préciser le sens des mots, identifier les évolution syntaxiques ou sémantiques, etc. Enfin, dans le cas des langues finno-ougriennes, il s'agit aussi de produire des corpus annotés pérennes pour des langues parfois gravement en danger. Le projet contribue donc à la documentation de ces langues, une entreprise urgente alors que chaque année des langues disparaissent, trop souvent sans laisser de trace.

Un projet en cours de démarrage est le projet ANR Democrat, porté par Frédéric Landragin, autour de l'analyse des chaînes de référence dans les textes. Ce projet fait l'objet de la section suivante.

D'autres collaborations concernent enfin des questions de recherche plus pointues pour lesquelles il est nécessaire de concevoir des développements propres. C'est notamment le cas des sciences sociales, de plus en plus souvent confrontées à de grandes collections de documents dont il faut tirer du sens sans en dénaturer le contenu, ni en offrir une vision par trop biaisée [Poibeau, 2014]. Le programme exploratoire PoliInformatics allait dans ce sens : il s'agissait, à partir d'une masse diversifiée de documents sur la crise financière de 2008-2009 aux Etats-Unis, de produire des analyses automatiques permettant à des experts du domaine d'identifier les principaux acteurs de la crise, leur rôles et surtout, dans la mesure où le corpus incluait essentiellement des textes post-crise (interviews de banquiers, de conseillers gouvernementaux, de membres d'organismes de régulation, etc. devant le Sénat américain par exemple), d'essayer d'identifier des points de vue consensuels ou contradictoires. Une tâche aujourd'hui relativement courante consiste à élaborer des réseaux d'acteurs, et de rendre compte graphiquement de leurs connexions sur le plan du contenu, des arguments et des opinions [Poibeau et Ruiz, 2015]. Ces travaux nous semblent intéressants car ils posent des questions d'analyse qui sont à la limite de l'état de l'art (analyse de l'argumentation, des prises de position, etc.). De très nombreuses annonces commerciales prétendent avoir résolu ce type de questions ou des problèmes similaires (comme l'analyse de l'opinion) mais ces travaux sont souvent très sommaires et reposant généralement sur des listes de mots clés prédéfinis sans tenir compte de l'application, du domaine ou du contexte. Ce type de recherche pose aussi des questions importantes sur le plan de l'éthique mais il nous semble important de les aborder pour contribuer à éclairer le débat public.

Les grands débats de société (comme les échanges sur le changement climatique), les élections locales ou nationales ou la production scientifique elle-même [Omodei *et al*., 2014] forment autant de données textuelles massives, peu structurées, que les outils automatiques peuvent aider à analyser. Il y a alors un réel apport des techniques de traitement automatique des langues : il ne s'agit plus de mettre en ligne des données, ni des les enrichir mais bien d'en extraire des informations nouvelles qu'il serait très difficile d'observer directement, sans outil numérique adapté. En ce sens, la recherche en Humanités numériques est apparue comme une chance à saisir pour le Lattice, à un moment où les outils de TAL semblent suffisamment matures pour être utilisables dans ce contexte nouveau, malgré leurs limites et leurs défauts. En retour, les Humanités numériques proposent un cadre permettant d'améliorer les outils existants et d'en concevoir de nouveaux, voire de concevoir des problématiques nouvelles.

Le projet DEMOCRAT :

Financé par l'ANR dans le cadre de l'appel à projets générique 2015, défi 8 « Sociétés innovantes, intégrantes et adaptative », le projet DEMOCRAT fait suite à un projet PEPS INS2I-INSHS (CNRS) intitulé MC4 (« Modélisation Contrastive et Computationnelle des Chaînes de Coréférence »), porté par Frédéric Landragin entre 2011 et 2013, qui a fédéré des chercheurs de trois unités de recherche travaillant tous sur les chaînes de référence : des chercheurs du Lattice, d'ICAR et de LiLPa. C'est dans ce cadre que les collaborations entre ces trois laboratoires se sont amorcées. Le projet MC4 avait pour objectif un double volet : linguistique descriptive, d'une part, et linguistique outillée et automatique, d'autre part. Il réunissait ainsi des spécialistes de linguistique du français contemporain et médiéval, reconnus pour leurs compétences dans le domaine de la linguistique référentielle, et des chercheurs en linguistique informatique, spécialisés également dans les questions de référence et de saillance référentielle. Les résultats principaux de ce projet, outre la mise en place de DEMOCRAT, sont d'une part la mise en œuvre d'un corpus annoté en chaînes de référence – corpus de taille modeste paru en juin 2015 sur la plateforme ORTOLANG – et d'autre part un ensemble d'articles de recherche regroupés dans un numéro thématique de la revue Langages (cf. bibliographie).

Objectifs :

DEMOCRAT vise à développer les recherches sur la langue et la structuration textuelle du français *via* l'analyse détaillée et contrastive des chaînes de référence (instanciations successives d'une même entité) dans un corpus diachronique de textes écrits entre le 9ème et le 21ème siècle, avec des genres textuels variés. Le projet mettra à disposition de la communauté scientifique : (i) un modèle intégré et discursif de la référence et de la composition des chaînes de référence ; (ii) un corpus annoté qui puisse servir de corpus de référence et de corpus d'apprentissage pour les campagnes d'évaluation internationales portant sur la coréférence ; (iii) un outil d'annotation, d'aide à l'annotation et de manipulation des données annotées, et (iv) un système de détection automatique des coréférences. Le corpus annoté manuellement en chaînes de référence aura une taille de 1 million de mots, soit environ 100 000 maillons annotés.

Résultats ou retombées attendues :

Dans notre société numérique, les corpus de textes s'avèrent essentiels pour les recherches scientifiques, la diffusion des connaissances et du patrimoine, la pérennisation et la standardisation des données. DEMOCRAT contribue aux humanités numériques en proposant un corpus numérique riche (varié et diachronique), pour la langue française, annoté en fonction d'analyses linguistiques relevant d'une dimension encore peu explorée, à la fois sémantique et pragmatique.

En apportant de nouvelles connaissances et données sur la langue, ce corpus et le modèle associé sont destinés à : (i) nourrir l'ensemble des applications de traitement automatique des langues – résumé automatique, traduction automatique, simplification de textes, fouille de texte, web sémantique, dialogue humain-machine, etc. – (ii) renforcer la place du français dans le monde *via* notamment l'intégration du français dans des défis scientifiques d'ampleur internationale (SemEval, CoNLL), et (iii) apporter à toutes les disciplines connexes à la linguistique, comme la didactique, la

psycholinguistique, l'enseignement du français et des langues, de nouvelles connaissances sur l'accès aux entités d'un texte, sur le fonctionnement des chaînes de référence et leur importance au niveau de la structuration et de la cohésion textuelle.

Si le corpus DEMOCRAT relève pleinement des humanités numériques, la conception d'un outil de détection automatique des chaînes de référence relève, elle, de l'intelligence artificielle – tout en s'appuyant sur le corpus. La notion de chaîne de référence est l'un des éléments clés de la cohésion et de la cohérence textuelles : comprendre qu'un texte parle en continu d'une entité humaine (Barack Obama), organisationnelle (l'ONU) ou abstraite (la justice) permet *ipso facto* d'en déterminer le thème central, et partant, d'en faciliter le traitement, dont la mémorisation, les modes de restitution de textes comme le résumé, la paraphrase, l'indexation et l'extraction d'informations, sans oublier la traduction. C'est dans une telle optique de détermination du thème central que les moteurs de recherche s'intéressent actuellement à la détection automatique de chaînes de référence. De manière plus originale, la teneur des chaînes de référence donne des indices sur les opinions ou prises de position du locuteur vis-à-vis de ses objets de discours, et aide ainsi les travaux centrés sur la détection d'opinions. En 2011, dans un article intitulé « The Winograd Schema Challenge » présenté lors d'une conférence d'intelligence artificielle, la détection de chaînes de références a été mise en avant comme tâche pouvant remplacer le célèbre test d'intelligence d'Alan Turing, *via* la notion de schéma de Winograd, notion regroupant 1 ou 2 coréférences dans un couple de phrases bien choisies. Les retombées attendues de DEMOCRAT au niveau du traitement automatique des langues relèvent donc pleinement de l'intelligence artificielle.

Références bibliographiques :

Adèle Désoyer, Frédéric Landragin, Isabelle Tellier, Anaïs Lefeuvre, Jean-Yves Antoine. Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues* n°55(2), 2014, pp. 97-121. <halshs-01153297>

Frédéric Landragin. Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus* n°10, 2011, pp. 61-80. <halshs-00658362>

Frédéric Landragin, Thierry Poibeau, Bernard Victorri. ANALEC: a New Tool for the Dynamic Annotation of Textual Data. *Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turquie, 2012, pp. 357-362. <halshs-00698971>

Frédéric Landragin, Catherine Schnedecker (Eds). Les chaînes de référence. *Langages* n°195 (septembre 2014), Armand Colin.

Frédérique Mélanie-Bécquet, Johan Ferguth, Katherine Gruel, Thierry Poibeau (2015). Archaeology in the Digital Age: From Paper to Databases. Actes de la Conférence Digital Humanities 2015, Sydney, Australie. <hal-01173964>

Elisa Omodei, Yufan Guo, Jean-Philippe Cointet, Thierry Poibeau (2014). Analyse discursive automatique du corpus ACL Anthology. *21ème conférence Traitement Automatique des Langues Naturelles*, Jun 2014, Marseille, France. 6 p., 2014. <hal-01056143>

Thierry Poibeau (2014). Le traitement automatique des langues pour les sciences sociales, quelques éléments de réflexion à partir d'expériences récentes. Revue *Réseaux* n°188 (2014/6).

Thierry Poibeau, Pablo Ruiz (2015). Generating Navigable Semantic Maps from Social Sciences Corpora. Actes de la Conférence Digital Humanities 2015, Sydney, Australie. <hal-01173963>

Prévost S. (2015) « Diachronie du français et linguistique de corpus : une approche quantitative renouvelée », *Langages*, 197 'La fréquence textuelle : un état des lieux', p. 23- 45.