Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Arthur AMALVY : arthur.amalvy@univ-avignon.fr

Titre : Traitement du langage naturel appliqué à la représentation de textes narratifs par réseaux de personnages

Mots-clés : réseaux de personnages, traitement du langage naturel, textes narratifs.

Title: *Natural Language Processing for the Representation of Narrative Texts through Character Networks*

Keywords: *character networks, natural language processing, narrative texts.*

Thèse de doctorat en informatique, laboratoire informatique d'Avignon, UFR sciences, technologies, santé, université d'Avignon, sous la direction de M. Vincent Labatut (MC, université d'Avignon, LIG) et M. Richard Dufour (Pr, Nantes Universités, LS2N). Thèse soutenue le 09/12/2024.

Jury : M. Vincent Labatut (MC, université d'Avignon, LIG, codirecteur), M. Richard Dufour (Pr, Nantes Universités, LS2N, codirecteur), Mme Claire Gardent (DR, CNRS, LORIA, rapporteuse, présidente), M. Christophe Cerisara (CR, CNRS, LORIA, rapporteur), M. David Bamman (Associate Professor, UC Berkeley School of Information, Berkeley, Californie, États-Unis, examinateur), Mme Farah Benamara (Pr, université Toulouse III – Paul Sabatier, IRIT, examinatrice).

Résumé : *Un réseau de personnages représente des personnages comme des sommets dans un graphe, et leurs relations comme les arêtes entre ces sommets. Dans le cas des œuvres littéraires, ils permettent de modéliser un récit entier en utilisant un seul objet mathématique. Leurs arêtes peuvent représenter différents types d'interactions : co-occurrence, conversation, action directe… De plus, l'évolution temporelle des relations peut être modélisée avec des réseaux dynamiques. Grâce à cette flexibilité, les*

réseaux de personnages ont été utilisés pour s'attaquer à plusieurs tâches, comme la classification de genre littéraire, la segmentation de récit, ou encore le résumé automatique. Extraire ces réseaux manuellement est cependant coûteux, et de nombreux chercheurs sont donc intéressés par l'automatisation de ce processus, qui nécessite la résolution de plusieurs tâches de traitement du langage naturel.

Dans cette thèse, nous présentons des contributions à ce processus d'extraction automatique dans le cas des romans, ainsi qu'à des applications des réseaux de personnages. Nous proposons Renard, un pipeline d'extraction modulaire sous licence libre. Nous l'utilisons pour étudier l'impact des erreurs de reconnaissance d'entités nommées (REN) et de résolution de coréférences sur la qualité des réseaux extraits. Nous observons que la performance des deux tâches est importante, et dépend fortement du roman étudié. Pour la résolution de coréférences, nous notons que l'impact dépend du type d'erreur.

En outre, nous identifions et contribuons à deux défis des systèmes d'extraction de réseaux de personnages. Le premier est le manque de données littéraires pour entraîner ces systèmes. Nous nous y attaquons d'une part en proposant un nouveau jeu de données littéraires couvrant la REN et la résolution d'alias et d'autre part en proposant d'utiliser une technique d'augmentation de données dans le cas de la REN interdomaines. Le second défi que nous identifions est la portée limitée des modèles à base de transformers*, qui peut être préjudiciable à la performance de certaines tâches. Nous proposons de récupérer du contexte pertinent au niveau du document pour atténuer le manque d'information induit par cette faible portée, et montrons que cela peut augmenter la performance de la tâche de REN.*

Enfin, nous présentons des contributions aux applications des réseaux de personnages dynamiques dans le cadre de deux études de cas. Premièrement, nous utilisons des réseaux modélisant différents types d'interactions dans une analyse de Lorenzaccio *d'Alfred de Musset. En utilisant la détection de communautés, nous identifions et étudions les différentes intrigues de la pièce. De plus, nous proposons une méthode automatique pour détecter des conspirations. Deuxièmement, nous proposons d'employer les réseaux de personnages pour résoudre la tâche d'alignement narratif sur trois adaptations du* Trône de fer *de George R. R. Martin (romans,* comics *et série télévisée). Nos résultats montrent que les méthodes basées sur les réseaux peuvent*

être meilleures que celles basées sur le texte, et que leur combinaison permet d'améliorer la performance. Nous mettons aussi en valeur l'importance de réaliser la tâche d'alignement sur des unités narratives commensurables.

URL où le mémoire peut être téléchargé :
https://theses.fr/s379155

Ousseynou J. M. GUEYE : gueyeousseynou@hotmail.fr

Titre : Personnalisation adaptative de problèmes mathématiques arithmétiques pour élèves de CM1-CM2 à l'aide de grands modèles de langue via ingénierie de prompt

Mots-clés : génération automatique, traitement automatique des langues, TAL, ingénierie de prompt, grands modèles de langue, LLM, enseignement personnalisé, linguistique de corpus, résolution de problèmes, technologie éducative.

Title: *Adaptive Personalization of Arithmetic Math Statements for Fourth and Fifth Grade Students Using Large Language Models through Prompt Engineering*

Keywords: *automatic generation, NLG, natural language processing, NLP, prompt engineering, large language models, LLM, personalized education, corpus linguistics, problem-solving, educational technology.*

Thèse de doctorat en sciences du langage et traitement automatique des langues, modèles, dynamiques, corpus, MoDyCo, UMR 7114, école doctorale connaissance, langage, modélisation, ED 139, université Paris Nanterre, sous la direction de Mme Iris Eshkol-Taravella (Pr, université Paris Nanterre). Thèse soutenue le 03/12/2024.

Jury : Mme Iris Eshkol-Taravella (Pr, université Paris Nanterre, directrice), M. Guillaume Desagulier (Pr, université Bordeaux Montaigne, CLIMAS, président), M. Didier Schwab (Pr, université Grenoble Alpes, GETALP-LIG, rapporteur), Mme Natalia Grabar (CR, HDR, CNRS, STL, Lille, rapporteuse), M. Patrick Paroubeck (IR, HDR, université Paris-Saclay, LISN, examinateur), M. Damien Nouvel (MC, institut national des langues et civilisations orientales, INALCO, ERTIM, examinateur).

Résumé : *Cette thèse en traitement automatique des langues (TAL) explore la personnalisation adaptative de problèmes arithmétiques pour les élèves de CM1-CM2 à l'aide de grands modèles de langue (LLM) via ingénierie de prompt. Elle se base sur des fondements théoriques en pédagogie, didactique, et linguistique.*

Cette recherche répond à un besoin croissant de mise à disposition d'outils éducatifs permettant d'améliorer la compréhension et l'engagement des élèves, notamment en mathématiques. En constituant un corpus spécialisé de problèmes mathématiques, ce travail analyse les caractéristiques linguistiques influençant la compréhension et la

résolution des problèmes. Ces analyses théoriques et méthodologiques ont ensuite été mises en pratique dans le développement de la plateforme Mathify101.

Mathify101est une plateforme web conçue pour générer des problèmes mathématiques personnalisés, adaptés aux besoins des élèves et des enseignants. Elle intègre une méthodologie rigoureuse d'ingénierie de prompt, permettant de personnaliser les problèmes tout en garantissant leur intégrité pédagogique.

Les expérimentations menées en milieu scolaire ont montré le potentiel de Mathify101à améliorer la motivation des élèves et leur maîtrise des concepts arithmétiques fondamentaux. De plus, l'évaluation automatique des problèmes générés a démontré leur conformité aux critères issus du corpus, validant ainsi l'efficacité du modèle proposé.

Les résultats de cette thèse soulignent l'importance des approches interdisciplinaires dans les technologies éducatives, en montrant comment les modèles de TAL peuvent être ajustés pour répondre à des objectifs pédagogiques spécifiques. Cette recherche contribue à l'évolution des technologies d'apprentissage adaptatif, en fournissant un cadre pour les développements futurs dans le domaine de l'éducation personnalisée.

URL où le mémoire peut être téléchargé :
https://theses.fr/s258241

Bingzhi LI : bingzhi2013@gmail.com

Titre : Étude des capacités abstractives de modèles de langue neuronaux

Mots-clés : traitement automatique des langues, modèles de langue neuronaux, interprétabilité, généralisation, abstraction linguistique, representation syntaxique, structures hiérarchiques, compositionnalité.

Title: *Study of the Abstraction Capabilities of Neural Language Models*

Keywords: *natural language processing, neural language models, linguistic abstraction, interpretability, generalizations, syntactic representation, hierarchical structures, compositionality.*

Thèse de doctorat en sciences du langage, laboratoire de linguistique formelle, LLF, UFR de linguistique, Université Paris Cité, sous la direction de M. Benoît Crabbé (Pr, Université Paris Cité) et M. Guillaume Wisniewski (MC, Université Paris Cité). Thèse soutenue le 28/11/2023.

Jury : M. Benoît Crabbé (Pr, Université Paris Cité, codirecteur), M. Guillaume Wisniewski (MC, Université Paris Cité, codirecteur), Mme Barbara Hemforth (DR, CNRS, présidente), M. Thierry Poibeau (DR, CNRS, ENS-PSL, rapporteur), M. François Yvon (DR, CNRS, rapporteur), Mme Dieuwke Hupkes (*research scientist*, Meta AI, examinatrice).

Résumé : *Traditional linguistic theories have long posited that human language competence is founded on innate structural properties and symbolic representations. However, transformer-based language models, which learn language representations from unannotated text, have excelled in various natural language processing tasks without explicitly modeling such linguistic priors. Their empirical success challenges these long-standing linguistic assumptions and also raises questions about the models' underlying mechanisms for linguistic competence. However, the black-box nature and complexity of these models, due to their numerous parameters, make it difficult to understand their internal workings. While research in this area is growing, the extent of their linguistic abstraction capabilities remains an open question. This thesis seeks to determine whether transformer models primarily rely on surface-level patterns for representing syntactic structures, or if they also implicitly capture more abstract rules. The study serves two main objectives: i) assessing the potential of an autoregressive transformer language model as an explanatory tool for human syntactic processing; ii) enhancing the model's interpretability. To achieve these goals, we assess the syntactic abstractions in transformer models on two levels: first, the ability to represent hierarchical structures, and second, the ability to compositionally generalize observed structures. We introduce an integrated linguistically informed analysis framework that consists of three interrelated layers: behavioral assessment through challenge sets, representational probing using linguistic probes, and functional analysis through causal intervention. Our analysis starts with assessing the model's performance on syntactic challenge sets to see how closely it mirrors human language behavior. Following this, we use linguistic probes and causal interventions to assess how well the model's internal representations align with established linguistic theories. Our findings reveal that transformers manage to represent hierarchical structures for nuanced syntactic generalizations. However, instead of relying on systematic compositional rules, they seem to lean more towards lexico-categorical abstraction and structural analogies. While this allows them to handle a sophisticated form of grammatical productivity for familiar structures, they encounter challenges with structures that require a systematic application of compositional rules. This study highlights both the promise and potential limitations of autoregressive transformer models as explanatory tools for human syntactic processing, and provides a methodological framework for its analysis and interpretability.*

URL où le mémoire peut être téléchargé :
https://theses.fr/2023UNIP7255

Aboubacar TUO : tuo.aboubacar@yahoo.fr

Titre : Extraction d'événements à partir de peu d'exemples par méta-apprentissage

Mots-clés : apprentissage de représentation, apprentissage frugal, extraction d'événements, méta-apprentissage, injection de connaissances.

Title: *Meta-learning for Few-shot Event Extraction*

Keywords: *representation learning, event extraction, few-shot learning, meta-learning, knowledge injection.*

Thèse de doctorat en informatique, laboratoire d'intégration des systèmes et des technologies, CEA-List, faculté des sciences d'Orsay, université Paris-Saclay, sous la direction de M. Olivier Ferret (DR, CEA list, LASTI). Thèse soutenue le 20/12/2023.

Jury : M. Olivier Ferret (DR, CEA list, LASTI, directeur), Mme Claire Nédellec (DR, INRAE, présidente), M. Vincent Claveau (CR, HDR, CNRS, IRISA, Rennes, rapporteur), M. Christophe Gravier (Pr, université Jean Monnet, Saint-Étienne, rapporteur), M. Matthieu Labeau (MC, institut polytechnique de Paris, examinateur).

Résumé : *L'extraction d'information est un champ de recherche dont l'objectif est d'identifier et extraire automatiquement des informations structurées, dans un domaine donné, à partir de données textuelles non structurées. Cette discipline trouve de nombreuses applications pratiques, de l'analyse automatique de documents médicaux à l'exploitation de flux d'actualités. Cependant, la mise en œuvre de telles extractions demande souvent des moyens humains importants pour l'élaboration de règles d'extraction ou encore pour la constitution de données annotées pour les systèmes utilisant de l'apprentissage automatique. Un des défis actuels dans le domaine de l'extraction d'information est donc de développer des méthodes permettant de réduire, dans la mesure du possible, les coûts et le temps de développement de ces systèmes.*

Ce travail de thèse se concentre sur l'exploration de l'extraction d'événements à travers l'utilisation du méta-apprentissage, une approche particulièrement adaptée à l'apprentissage à partir de peu de données. Cette approche permet à un modèle d'acquérir des connaissances généralisables à partir d'une variété de tâches. Nous avons redéfini la tâche d'extraction d'événements dans cette perspective, cherchant à développer des systèmes capables de s'adapter rapidement à de nouveaux contextes d'extraction avec un faible volume de données d'entraînement.

Dans un premier temps, nous avons proposé des méthodes visant à améliorer la détection des déclencheurs événementiels en développant des représentations plus robustes pour cette tâche, à travers la combinaison des représentations issues des couches cachées d'un modèle de langue. Cette approche permet de capturer des informations complémentaires à différents niveaux d'abstraction linguistique. Ensuite, nous avons abordé le défi spécifique posé par la classe « NULLE » (absence d'événement) dans ce cadre, un aspect souvent négligé, mais crucial pour la performance des systèmes,

en particulier dans un contexte d'extraction d'événements où cette classe est prédominante. Enfin, nous avons évalué l'effectivité de nos propositions dans le contexte global de l'extraction d'événements en les étendant à l'extraction des arguments des événements, démontrant ainsi la généralisation possible de notre approche à l'ensemble de la chaîne de traitement. Les résultats expérimentaux sur plusieurs jeux de données de référence confirment la pertinence de notre approche pour améliorer les performances des systèmes d'extraction d'événements dans un contexte de faibles ressources.

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04382185
