PROPOSITION DE SUJET DE THÈSE

CONTRATS DOCTORAUX
2021–2024

Appel ciblé : Contrat doctoral fléché FR Agorantic

Directeur de thèse : Richard Dufour email : richard.dufour@univ-avignon.fr
Laboratoire Informatique d'Avignon

Co-directeur de thèse : Guillaume Marrel email : guillaume.marrel@univ-avignon.fr
Laboratoire Biens, Normes, Contrats

Co-encadrant : Vincent Labatut email : vincent.labatut@univ-avignon.fr
Laboratoire Informatique d'Avignon

Titre en français : Détection d'évènements et extraction de réseaux d'interactions à partir de notices biographiques et d'articles de presse

Titre en anglais : Event detection and interaction network extraction based on bibliographical notes and newspaper articles

Résumé : Cette thèse vise d'abord à tirer parti du développement récent des approches neuronales utilisant l'apprentissage profond, afin de résoudre trois tâches de traitement du langage naturel sur des textes en français. Elle a ensuite pour objectif d'exploiter l'information ainsi extraite pour construire des graphes d'interactions sociales représentant les relations entre les personnes mentionnées par les textes. Enfin, son dernier volet concerne l'application des ces outils à un corpus de pages Wikipédia et d'articles de presse décrivant l'activité de personnes publiques, dans le but de répondre à des problématiques de science politique.

Mots-clés : Traitement automatique du langage naturel, Analyse de graphes, Représentation politique

1 Présentation détaillée du sujet 1
1.1 Contexte et enjeux . . . 1
1.2 Objectifs . . . 2
1.3 Méthode . . . 3
1.4 Organisation . . . 5

2 Profil du candidat ou de la candidate 6

3 Opportunités de mobilité 6

4 Références bibliographiques 6

1 Présentation détaillée du sujet

1.1 Contexte et enjeux

L'étude d'un groupe social et de son fonctionnement passe nécessairement par celle des membres qui le composent, ainsi que des relations qui existent entre eux. En fonction de l'objet de l'analyse, les relations unissant sociologiquement les membres d'un groupe social peuvent être constituées de différentes manières : par les interactions entre les différents acteurs, par leur participation à des

évènements, par leur appartenance à des structures ou encore par leur adhésion à des opinions spécifiques. Cette étude requiert d'identifier et de caractériser ces relations, ce qui peut être réalisé sur la base de documents publics exposant l'activité du groupe social et des interactions qui le structurent. L'approche peut s'appliquer à tous les groupes sociaux, historiques ou contemporains, pour lesquels on dispose de documents décrivant leur activité. Elle est donc particulièrement pertinente dans le cas des personnalités publiques du monde politique.

Les relations unissant les membres d'un groupe social peuvent être représentées sous la forme d'un graphe, au sens mathématique du terme, c'est-à-dire un ensemble de sommets représentant les individus, et un ensemble d'arêtes correspondant aux relations entre eux. L'utilité d'une telle représentation réside non seulement dans sa forme condensée permettant une visualisation relativement intuitive, mais également dans l'exploitation de ses propriétés topologiques à des fins d'analyse quantitative [21, 17], par exemple via l'identification d'acteurs centraux [5] ou la détection de communautés [4].

La source d'information la plus fournie et la plus simple à traiter pour en extraire ce type de graphe est à n'en pas douter la forme écrite (par opposition aux documents multimédias : vidéos, images, sons). Cependant, même à partir de documents écrits, la compilation et la caractérisation manuelle de ces relations dans le but de constituer un graphe de relations exploitable représente un travail qui peut être long, difficile, et sujet aux erreurs, suivant la nature et la taille du corpus utilisé. De plus, l'ampleur d'une telle tâche force parfois les annotateurs à se résoudre à des simplifications qui peuvent altérer le réseau final ou diminuer la pertinence de l'information extraite.

L'automatisation de ce travail d'extraction constituerait donc une réelle aide à la recherche en sciences sociales, et en science politique en particulier. Le gain de temps obtenu permettrait d'appréhender des quantités de données difficiles, voire impossibles, à traiter manuellement, et ce avec un niveau de fiabilité constant. Utiliser une grande quantité de données pour réaliser une analyse sociologique n'est pas pour autant un gage de qualité en soi, car on pourrait arguer que de telles méthodes automatiques ne peuvent déceler dans le texte que ce que son auteur a bien voulu y mettre. Cependant, aucune étude d'ampleur n'ayant été réalisée jusqu'à présent sur une telle quantité de données, un outil automatique permettrait déjà d'obtenir un aperçu objectif des informations décrites dans ce type de corpus, et, en creux, de celles qui en sont absentes. Enfin, soulignons que ce type d'outil ne peut pas se substituer à l'expertise du ou de la chercheuse. Il vise plutôt à lui donner accès à des informations jusque là inaccessibles, sur la base desquelles il ou elle peut ensuite échafauder son analyse.

L'extraction automatique de ces informations à partir d'un document textuel relève d'un sous-domaine de l'informatique appelé *Traitement Automatique du Langage Naturel* (TALN), et nécessite la résolution d'un certain nombre de tâches bien définies. Cependant, jusqu'à récemment, les méthodes automatiques existantes ne permettaient pas de résoudre ces tâches de façon satisfaisante, en particulier sur des textes français, langue pour laquelle les ressources linguistiques (corpus annotés, modèles de langages) sont moins nombreuses que pour l'anglais. Mais l'avènement récent des techniques d'apprentissage profond a radicalement changé la donne, et il semble aujourd'hui tout à fait envisageable de mettre au point des outils capables de réaliser les tâches de TALN requises par l'extraction de graphes sociaux.

1.2 Objectifs

Cette thèse se place à l'interface de l'informatique et de la science politique, et comporte trois objectifs principaux : deux d'ordre méthodologique et un d'ordre applicatif.

Le **premier objectif** est de construire et mettre en oeuvre des outils de TALN modernes permettant d'identifier les acteurs mentionnés dans un corpus de textes en langue française, mais aussi d'extraire et de caractériser leurs relations, et enfin de repérer les évènements dans lesquels ils sont impliqués.

Le **deuxième objectif** est de définir des méthodes automatiques de construction de graphes sociaux à partir des acteurs, interactions et évènements détectés via les outils de TALN. Son corollaire est l'élaboration et la mise en oeuvres de méthodes de visualisation et d'analyse adaptées aux graphes produits.

Le **troisième objectif** vise à valider les outils automatiques issus des deux premiers points. Il s'agit d'appliquer les outils de TALN et d'extraction de graphes décrits ci-dessus à un corpus de textes biographiques et journalistiques décrivant l'activité de personnalités politiques publiques, afin de

répondre à des problématiques de science politique.

1.3 Méthode

Les aspects méthodologiques sont clairement séparés entre TALN, extraction de graphes, et validation sur les données de science politique.

Tâches de TALN La première étape du traitement vise à résoudre les tâches de TALN. Nous en avons dénombré trois. Tout d'abord, il est nécessaire d'identifier les entités d'intérêt apparaissant dans le texte, une tâche appelée *Reconnaissance d'Entités Nommées*. Exprimé simplement, il s'agit de repérer les noms propres correspondant à des personnes, lieux, objets, etc. La deuxième tâche est appelée *Résolution de Co-références* et a pour but d'identifier dans le texte les constructions anaphoriques (notamment les pronoms et syntagmes nominaux) faisant référence aux entités détectées sans les mentionner explicitement. Enfin, la troisième tâche est l'*Extraction d'Évènements*, dont le but est de déterminer *qui* a fait *quoi*, *où*, *quand*, et *avec* qui, *sur* qui/quoi.

Ces dernières années, les approches neuronales, et notamment les réseaux de neurones mettant en jeu les techniques d'apprentissage profond, ont révolutionné plusieurs domaines de l'informatique, dont le TALN. On peut notamment citer les réseaux neuronaux convolutifs et les réseaux neuronaux récurrents [7], qui permettent de tirer parti de la nature séquentielle du texte, ou bien les auto-encodeurs [18] capables d'apprendre par eux mêmes une représentation compacte et pertinente des mots utilisable par des outils génériques. Plus récemment, les modèles de langage pré-entraînés permettant la vectorisation contextuelle de mots tels que BERT [3] ou XLNET [24] ont considérablement fait progresser l'état de l'art dans de nombreuses tâches du domaine.

La première étape de la thèse sera d'effectuer l'état de l'art des méthodes existantes pour les trois tâches identifiées. Ce travail sera réalisé non seulement d'un point de vue théorique, dans le but de comprendre les principes des méthodes proposées, mais aussi d'un point de vue pratique, en recensant lesquelles parmi elles ont été mises en application sur le français, et en considérant leur niveau de performance.

La détection d'entités nommées est une tâche pour laquelle un travail conséquent a déjà été publié, à la fois via des méthodes classiques [8] et neuronales [22]. Dans une moindre mesure, cela vaut également pour la résolution de co-références [10]. Ces deux tâches sont toutefois loin d'être complètement résolues. L'une des stratégies que nous envisageons pour les traiter est d'utiliser des méthodes d'apprentissage par transfert [19], qui consistent à entraîner un modèle sur des données différentes du corpus ciblé mais accessibles en grande quantité, puis d'affiner l'apprentissage sur une petite quantité des données ciblées. Un défi supplémentaire réside dans le fait que ces méthodes par transfert se basent généralement sur des modèles pré-entraînés sur des corpus en anglais. Notre travail se portant sur le français, il conviendra d'utiliser les modèles pré-entraînés sur cette langue, tels que CamemBERT [16], ou de développer une approche multilingue.

Pour la tâche d'extraction d'évènements, si des approches classiques existent [9], l'utilisation d'approches neuronales est beaucoup plus récente [20, 23], et moins explorée. De plus, la notion d'évènement peut se définir de nombreuses façon différentes, et la nature de cette définition affecte significativement les performances de ces méthodes. Dans le cadre de cette thèse, notre objectif est d'identifier des évènements spatiaux-temporels, permettant non seulement de mettre en relation des entités représentant des acteurs et d'identifier la nature de l'action qui les relie, mais également de contextualiser cette action en identifiant le lieu et la date à laquelle elle se produit, ainsi que, le cas échéant, les objets mis en oeuvre (par exemple deux auteurs écrivant un livre ensemble). L'une des pistes que nous envisageons est de nous baser sur les architectures neuronales existantes et d'en proposer des adaptations spécifiques à notre vision de cette tâche. Les autres pistes sont l'hybridation de méthodes traditionnelles et neuronales, ainsi que le développement de méthodes neuronales capables de traiter les trois tâches de TALN simultanément.

Construction du graphe La deuxième étape du traitement est l'extraction de graphes à partir des résultats fournis par les tâches de TALN : entités, mentions, évènements. Techniquement, plusieurs formes de graphes différents peuvent être ciblées, comme en témoigne la littérature existant sur l'extraction de réseaux de personnages fictionnels [11], un problème présentant de nombreuses

similarités avec l'objet de cette thèse, et sur lequel certains membres de l'équipe d'encadrement travaille déjà.

Une première approche consistera à transposer et appliquer des méthodes déjà développées dans ce contexte. Celles-ci permettent de détecter des relations **explicites** entre les acteurs identifiés dans le texte. Pour être bref, cela peut se faire de façon *approximative* en assimilant une co-occurrence d'acteurs (deux acteurs apparaissent au même endroit dans le texte) à une relation. Dans notre cas, l'information issue de la reconnaissance d'entités nommées et de la résolution de co-références est suffisante pour mettre en oeuvre cette approche. Une approche plus *précise* mais plus exigeante en termes de TALN consiste à relever une relation entre des acteurs seulement si elle est exprimée sous la forme d'une action impliquant les acteurs (un verbe a pour sujets ou objets les acteurs en question). Cette méthode peut être mise en oeuvre sur la base des évènements identifiés par notre troisième tâche de TALN.

Mais notre contexte applicatif nous fournit des informations plus complètes que celles extraites d'oeuvres de fiction, de point de vue de la contextualisation spatio-temporelle des évènements. En effet, dans un article journalistique ou une notice biographique, les évènements sont souvent datés et/ou localisés. Il est donc possible de détecter des relations **implicites** entre les acteurs, un objectif qui n'a pour ainsi dire pas encore été exploré dans la littérature. Prenons le cas de deux évènements possiblement mentionnés dans deux textes différents, et impliquant deux acteurs différents, mais dont le contexte est identique. Notre hypothèse est qu'il est possible, dans une certaine mesure, de faire l'hypothèse que ces deux acteurs entretiennent une relation d'une certaine forme, même si celle-ci n'est mentionnée explicitement à aucun moment dans le texte. Par exemple, deux personnes dont les biographies respectives indiquent qu'elles sont toutes les deux diplômées d'une même école et appartiennent à la même promotion se connaissent très certainement, même si leurs biographies ne l'indiquent pas.

La détection de ces relations implicites repose sur la détection d'évènements réalisée à l'étape de TALN. Le défi majeur consiste ici à unifier les évènements identifiés, c'est à dire à déterminer lesquels sont similaires, et dans quelle mesure, afin de relier les sommets concernés. La première méthode à mettre en oeuvre est algorithmique : il s'agit de développer manuellement une mesure de similarité entre évènements, qui nous permettra d'obtenir une performance de référence sur cette tâche. Il faudra pour cela utiliser la sémantique correspondant aux termes du texte décrivant l'action associée à l'évènement, ainsi qu'à son contexte. En effet, cette sémantique nous permettra d'effectuer des inférences, et ainsi de déterminer par exemple que le termes *Bordeaux* et *CUB*[1] ne sont pas exactement identiques, mais de sens suffisamment similaire pour en conclure une proximité géographique. Ceci implique d'utiliser les grandes bases de connaissances disponibles publiquement en ligne, telles que WikiData[2]. La seconde méthode que nous voulons explorer consiste à utiliser des approches neuronales pour apprendre une mesure de similarité entre évènements, directement à partir des données. Cette approche a l'avantage de ne pas reposer sur des pré-supposés pour définir ce que sont deux évènements similaires. Cependant, elle implique d'identifier un corpus annoté approprié.

Une fois les graphes construits, que ce soit via l'identification de relations explicites ou implicites, il sera nécessaire de les valider, afin d'estimer la fiabilité de notre approche. Ceci implique non seulement de constituer une *vérité terrain* en extrayant manuellement des graphes similaires à titre de référence, mais également d'appliquer une méthode de comparaison de graphes. La littérature est riche sur ce point, et présente un grand nombre de méthodes basées sur différents principes de théorie des graphes [1]. Nous comptons également explorer les approches neuronales pour traiter ce point [13], qui permettent d'apprendre la mesure de similarité à partir des données, au lieu de la définir *a priori*. Sur cet aspect, le candidat pourrait bénéficier du travail déjà réalisé dans le cadre de la thèse de Noé Cécillon sur les méthodes de plongements de graphes [2], qui se déroule actuellement au LIA sous la direction de Richard Dufour et Vincent Labatut.

Validation expérimentale Nos outils ont clairement vocation à être utilisés dans un contexte de sciences humaines et sociales, puisqu'ils ciblent l'étude des interactions sociales. La dimension interdisciplinaire apporte à l'informatique les enjeux socio-historiques, les éléments de contextualisation et d'évaluation pour tester les outils d'extraction d'information et de détection de réseaux sociaux.

1. Communauté urbaine de Bordeaux
2. https://www.wikidata.org/

Elle fournit aux politistes l'occasion d'équiper leurs recherches d'instruments d'une fouille de données textuelles systématisée et d'inventer alors de nouveaux corpus de données pour l'analyse des systèmes complexes que sont les réseaux d'acteurs du monde politique.

Une fois les étapes de TALN et d'extraction de graphes validées, nous voulons appliquer nos outils à des corpus de textes relatifs à l'activité du personnel politique français, afin d'étudier et de valider la pertinence de notre approche. Cela se fera en tentant de répondre à plusieurs questions relevant de la science politique. Cette partie du travail se place directement dans la continuité de la collaboration entre Guillaume Marrel et Vincent Labatut dans le cadre d'Agorantic, qui inclut notamment le projet RÉSOPO financé par Agorantic en 2015 et le travail mené sur la visibilité en ligne du personnel politique, qui avait abouti à deux communications [15, 12] et une publication [14].

L'application se concentrera sur l'étude du réseau de pouvoir d'un·e élu·e présidant le conseil d'une région importante. L'objectif est d'analyser la structure du réseau politique d'un leader régional, sa dimension, les ressources échangées, leurs usages [6]. Pour ce faire, nous constituerons d'abord un corpus qui soit le plus complet possible, basé sur plusieurs sources. La première sera Wikipedia, l'encyclopédie universelle et multilingue en ligne lancée en 2001, qui constitue une collection de sources singulière, coopérative, presque illimitée et sans cesse renouvelée. La collecte sera effectuée sur le principe du *Web-crawl*, en partant de la notice biographique personnelle de la personne publique ciblée et en progressant récursivement vers d'autres articles personnels ou institutionnels auxquels les hyperliens de la première renvoient. Cet ensemble de documents sera complété de sources secondaires issues du Web, sur la base de résultats renvoyés par des moteurs de recherche de type Google. D'après notre expérience sur une tâche similaire [15, 12, 14], ce type de recherche produit essentiellement des articles de presse et des billets de blog, constituant un ensemble hétérogène de documents.

Nous appliquerons alors nos outils afin d'extraire plusieurs graphes d'interactions, en faisant varier les méthodes et paramètres. Le but de cette opération est d'explorer leur effet sur le résultat final, et d'identifier les paramétrages les plus pertinents dans ce contexte applicatif. La qualité du réseau de pouvoir extrait sera évaluée grâce à l'expertise de Guillaume Marrel dans le domaine. L'apport de la méthode évènementielle que nous proposons est notamment d'extraire de l'information non seulement de chaque texte pris individuellement, en y détectant des entités et les évènements auxquels elles participent, mais aussi de la collection elle-même, à travers la comparaison d'évènements identifiés dans des textes distincts. L'un des principaux enjeux de cette partie applicative sera d'étudier ces relations implicites, et d'évaluer leur importance (et donc l'apport de notre méthode) du point de vue de la science politique.

1.4 Organisation

Thématiquement, le projet s'intègre dans plusieurs axes de la FR Agorantic :

- Axe 1 *Méthodologies et Interdisciplinarité* : la plus grande partie du travail de cette thèse est d'ordre méthodologique, et vise à équiper les SHS d'outils de fouille de texte applicables dans un grand nombre de domaines traités dans le cadre d'Agorantic.
- Axe 3 *Politique(s), transparence et éthique* : la partie applicative de la thèse est consacrée au traitement de problématiques de science politique..
- Axe 5 *Structuration et exploitation de corpus* : la conduite de la thèse requiert de constituer et d'annoter plusieurs types de corpus textuels qui seront ensuite mis à disposition de la FR. L'outil élaboré pourra être utilisé par d'autres chercheurs pour constituer d'autres corpus.

Les principales étapes du travail de thèse proposé suivent la décomposition détaillée en Section 1.3 :

1. **Identification automatique des acteurs et de leurs mentions.** Évaluation des méthodes existantes, et développement de méthodes spécialisées.
2. **Extraction automatique d'évènements.** Mise en place et évaluation d'une première version basée sur les approches existantes, puis amélioration progressive basée sur les métriques d'évaluation.
3. **Construction de graphes d'interactions.** Mise en place de l'approche algorithmique, puis de l'approche à base d'apprentissage automatique. Évaluation, comparaison.
4. **Validation expérimentale.** Mise en oeuvre de nos outils pour répondre aux problématiques de science politique identifiées en Section 1.3.

2 Profil du candidat ou de la candidate

La candidate ou le candidat devra posséder les compétences suivantes :

- Détenteur d'un diplôme de master en informatique ou d'ingénieur en informatique.
- Maîtrise du français et bon niveau d'anglais, à l'oral et à l'écrit.
- Maîtrise des langages de programmation C++, Java, et/ou Python.
- A suivi un cours de traitement automatique du langage naturel écrit et/ou possède une expérience pratique dans une tâche de TALN écrit : détection d'entités nommées, résolution de co-références, extraction d'évènements, etc.
- A suivi un cours sur l'apprentissage automatique profond, et/ou possède une expérience pratique mettant en oeuvre ce type de méthodes.

3 Opportunités de mobilité

Nous envisageons de mettre en place un stage doctoral sur la base d'un financement Hubert Curien STAR (Science and Technology Amical Relationship). Il s'agit de projets binationaux France-Corée du Sud, impliquant une équipe dans chacun des deux pays, et permettant à certains de leurs membres d'effectuer des missions chez l'autre. En l'occurrence, nos contacts en Corée sont O-Joun Lee (Pohang University of Science and Technology) et Jason J. Jung (Chung-Ang University). Ils travaillent sur l'extraction de réseaux de personnages dans des scripts, et leur caractérisation par des méthodes neuronales afin d'élaborer des plongements permettant de réaliser différentes tâches de recherche d'information sur ces représentation vectorielles. Leur travail et les aspects méthodologiques de cette thèse sont donc très complémentaires et susceptibles de se compléter.

4 Références bibliographiques

[1] M. Baur et M. Benkert. "Network comparison". In : *Network Analysis : Methodological Foundations*. T. 3418. Lecture Notes in Computer Science. 2005, p. 318-340. doi : 10.1007/978-3-540-31955-9_12.

[2] N. Cécillon et al. "Graph embeddings for Abusive Language Detection". In : *Springer Nature Computer Science* 2 (2021), p. 37. doi : 10.1007/s42979-020-00413-7.

[3] J. Devlin et al. In : *Conference of the North American Chapter of the Association for Computational Linguistics*. 2019, p. 4171-4186. doi : 10.18653/v1/n19-1423.

[4] S. Fortunato. "Community detection in graphs". In : *Physics Reports* 486.3-5 (2010), p. 75-174. doi : 10.1016/j.physrep.2009.11.002.

[5] L. C. Freeman. "Centrality in Social Networks I : Conceptual Clarification". In : *Social Networks* 1.3 (1978), p. 215-239. doi : 10.1016/0378-8733(78)90021-7.

[6] G. Garrote. "Entre franchissement et enfermement : pluralité et variabilité de configuration des réseaux notabiliaires territorialisés". In : *Les réseaux dans le temps et dans l'espace*. 2013, p. 143-161. url : https://groupefmr.hypotheses.org/2890.

[7] I. Goodfellow, Y. Bengio et A. Courville. *Deep learning - Adaptive computation and machine learning*. MIT Press, 2016. url : https://www.deeplearningbook.org/.

[8] A. Goyal, V. Gupta et M. Kumar. "Recent Named Entity Recognition and Classification techniques : A systematic review". In : *Computer Science Review* 29 (2018), p. 21-43. doi : 10.1016/j.cosrev.2018.06.001.

[9] F. Hogenboom et al. "An Overview of Event Extraction from Text". In : *Detection, Representation, and Exploitation of Events in the Semantic Web*. T. 779. CEUR Workshop Proceedings. 2011, p. 48-57. url : http://ceur-ws.org/Vol-779/.

[10] M. Joshi et al. "BERT for Coreference Resolution : Baselines and Analysis". In : *Conference on Empirical Methods in Natural Language Processing / 9th International Joint Conference on Natural Language Processing*. 2019, p. 5803-5808. doi : 10.18653/v1/D19-1588.

[11] V. Labatut et X. Bost. "Extraction and Analysis of Fictional Character Networks : A Survey". In : *ACM Computing Surveys* 52.5 (2019), p. 89. doi : 10.1145/3344548.

[12] V. Labatut et G. Marrel. "La visibilité politique en ligne : Contribution à la mesure de l'e-reputation politique d'un maire urbain". In : *Big Data et visibilité en ligne : Un enjeu pluridisciplinaire de l'économie numérique*. 2017. url : https://univ-droit.fr/actualites-de-la-recherche/44-manifestation-scientifique/25026-big-data-et-visibilite-en-ligne-un-enjeu-pluridisciplinaire-de-l-economie-numerique.

[13] G. Ma et al. "Deep Graph Similarity Learning : A Survey". In : *arXiv* cs.LG (2019), p. 1912.11615. url : https://arxiv.org/abs/1912.11615.

[14] G. Marrel et V. Labatut. "La visibilité politique en ligne de la maire de Paris - Contribution à la mesure de l'écho Web-médiatique d'Anne Hidalgo". In : *Big Data et visibilité en ligne - Un enjeu pluridisciplinaire de l'économie numérique*. Sous la dir. de Christophe Alcantara, Francine Charest et Serge Agostinelli. Presses des Mines, 2018, p. 271-286. url : https://www.pressesdesmines.com/produit/big-data-et-visibilite-en-ligne/.

[15] G. Marrel, V. Labatut et M. El Bèze. "Le Web comme miroir du travail politique quotidien ? Reconstituer l'écho médiatique en ligne des événements d'un agenda d'élu". In : *13ème Congrès de l'Association Française de Science Politique (AFSP)*. Aix-en-Provence, FR, 2015, p. 25. url : http://www.congres-afsp.fr/st/st7/st7marrellabatutelbeze.pdf.

[16] L. Martin et al. "CamemBERT : a Tasty French Language Model". In : *58th Annual Meeting of the Association for Computational Linguistics*. 2020, p. 7203-7219. doi : 10.18653/v1/2020.acl-main.645.

[17] P. Mercklé. *La sociologie des réseaux sociaux*. Repères. La Découverte, 2016. doi : 10.3917/dec.merck.2011.01.

[18] T. Mikolov et al. "Distributed representations of words and phrases and their compositionality". In : *26th International Conference on Neural Information Processing Systems*. 2013, p. 3111-3119. url : https://dl.acm.org/doi/10.5555/2999792.2999959.

[19] S. J. Pan et Q. Yang. "A Survey on Transfer Learning". In : *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), p. 1345-1359. doi : 10.1109/tkde.2009.191.

[20] D. Wadden et al. "Entity, Relation, and Event Extraction with Contextualized Span Representations". In : *Conference on Empirical Methods in Natural Language Processing / 9th International Joint Conference on Natural Language Processing*. 2019, p. 5784-5789. doi : 10.18653/v1/d19-1585.

[21] S. Wasserman et K. Faust. *Social Network Analysis : Methods and Applications*. T. 8. Structural Analysis in the Social Sciences. Cambridge, UK : Cambridge University Press, 1994. url : http://www.cambridge.org/zw/academic/subjects/sociology/sociology-general-interest/social-network-analysis-methods-and-applications.

[22] I. Yamada et al. "LUKE : Deep Contextualized Entity Representations with Entity-aware Self-attention". In : *Conference on Empirical Methods in Natural Language Processing*. 2020, p. 6442-6454. doi : 10.18653/v1/2020.emnlp-main.523.

[23] S. Yang et al. "Exploring Pre-trained Language Models for Event Extraction and Generation". In : *57th Annual Meeting of the Association for Computational Linguistics*. 2019, p. 5284-5294. doi : 10.18653/v1/p19-1522.

[24] Z. Yang et al. "XLNet : Generalized Autoregressive Pretraining for Language Understanding". In : *33rd Conference on Neural Information Processing Systems*. 2019. url : https://papers.nips.cc/paper/2019.