Curriculum Vitae

Ahmed Hamdi

ahmed.hamdi@univ-lr.fr
Laboratoire L3i,
Faculté des Sciences et Technologies Avenue Michel Crépeau
17042, La Rochelle
https://pageperso.univ-lr.fr/ahmed.hamdi/

Parcours universitaire

2011-2015
Diplôme : Doctorat
Discipline : Sciences de langage
Lieu : Aix-Marseille université
Mention : Très honorable

2010-2011
Diplôme : Master
Discipline : Sciences de langage
Spécialité : Traitement automatique des langues
Lieu : Aix-Marseille université
Mention : Bien

2007-2010
Diplôme : Ingénieur
Discipline : Informatique
Lieu : École nationale de sciences de l'informatique (Tunisie)
Mention : Assez bien

2005-2007
Diplôme : Premier cycle
Discipline : Mathématique-Physique
Lieu : Institut préparatoire des études d'ingénieur d'El Monastir (Tunisie)
Mention : Assez bien

Expériences professionnelles

2017-Aujourd'hui
Poste : Post-doctorant et vaccataire
Lieu : La Rochelle Université
Unité de recherche : Laboratoire Informatique, Image et Interaction (L3i)

2014-2016
Poste : Attaché temporaire d'enseignement et de recherche (ATER)
Lieu : Aix-Marseille Université
Unité de recherche : Laboratoire d'informatique fondamentale (LIF)

2011-2014
Poste : Doctorant et moniteur
Lieu : Aix-Marseille Université
Unité de recherche : Laboratoire d'informatique fondamentale (LIF)

Activités de recherche

2018-Aujourd'hui **Poste :** Post-doctorant
Unité de recherche : Laboratoire Informatique, Image et Interaction (L3i)
Lieu : La Rochelle Université
Projets : NewsEye, Labcom-IDEAS
Superviseurs : Antoine Doucet, Mickaël Coustaty
Contributions :

- **Reconnaissance et désambiguïsation d'entités nommées à partir de la presse ancienne** : La reconnaissance d'entités nommées (NER) est la tâche qui consiste à localiser et à catégoriser les concepts importants d'un texte donné dans un ensemble de classes prédéfinies (personne, lieu, organisation, évènement...). Cette tâche appliquée sur des documents historiques est confronté à trois défis principaux: le bruit du texte (documents usés), une dynamique du langage (évolution des standards orthographiques) et le manque de ressources annotées. Pour faire face à ces problèmes, j'ai participé au développement de techniques de reconnaissance d'entités nommées basées sur l'apprentissage par transfert et l'utilisation de transformateurs supplémentaires. Ces techniques ont amélioré les résultats de NER par rapport aux systèmes existants sur des journaux anciens. Ce travail a été récompensé par plusieurs publications scientifiques dans des conférences bien classées. Dans le cadre de ce travail, notre équipe a aussi remporté la compétition HIPE-CLEF 2020. En outre, j'ai participé à l'encadrement d'un étudiant en L3 de l'université de Hanoï appuyé par une publication dans la conférence ICADL (Core A).
- **Détection d'opinions envers les entités nommées** La détection d'opinions a pour objectif d'identifier l'opinion d'un auteur (c'est-à-dire en faveur, contre ou neutre) envers une entité cible (par exemple, une personne, une organisation, etc.). Dans le contexte du projet NewsEye, nous nous sommes concentrés sur la détection d'opinions dans les documents historiques. Nos méthodes de détection d'opinions jugent si un article de presse est posétif, négatif ou neutre envers une entité nommée donnée mentionné dans le texte. En d'autres termes, la tâche peut être considérée comme un problème de classification qui catégorise deux morceaux de texte, à savoir l'article et l'entité cible, en trois classes (positive, négative et neutre). Afin de proposer une étude exhaustive, nous avons défini trois méthodes:
  - méthode basée sur un lexique de sentiments: l'opinion de l'auteur envers une entité nommée est définie par le scores des mots qui l'entourent.
  - méthode basée sur des classifieurs: l'opinion de l'auteur envers une entité nommée est définie à l'aide d'un vote majoritaires sur les résultats des classifieurs entraînés sur un corpus annoté.
  - méthode basée sur l'apprentissage profond: similaire à la méthode

précédente sauf que la classification est réalisée à l'aide de BERT.

- **Extraction d'informations à partir de documents administratifs** : ce travail consiste à extraire des champs clés à partir de documents administratifs (factures, contrats, etc.) en utilisant deux méthodes différentes basées sur l'étiquetage séquentiel.

2017-2018

Poste : Post-doctorant
Unité de recherche : Laboratoire Informatique, Image et Interaction (L3i)
Lieu : La Rochelle Université
Projet : SecurDoc
Superviseurs : Mickaël Coustaty
Contributions :

- **Segmentation de flux de documents** : Les grandes entreprises traitent chaque jour un énorme flux de documents. Chaque document est composé d'une ou plusieurs pages ayant une relation logique et structurelle entre elles. Quand le flux est numérisé, il perd sa structure initiale et la numérisation génère un seul document qui fusionnent tous les documents. Afin de conserver les informations de début et de fin de chaque document, les techniques les plus répandues consiste à disposer d'un séparateur explicite à l'instar de pages blanches ou de tampons afin de délimiter les documents. Ces techniques, dès lors que le séparateur est détecté, sont fiables et rapides, cependant elles nécessitent une action manuelle énorme et le risque de l'erreur augmente surtout que le nombre de documents à traiter par jour est conséquent (environ 8000 pages/jour). C'est pour cette raison que des techniques de reconstitution automatique de flux de documents sont de plus en plus demandées.
  Mon activité comme étant post-doctorant à La Rochelle université avait pour objectif de proposer une méthode générique pour structurer automatiquement un flux documentaire. Dans ce cadre, j'ai utilisé un système existant proposé par des développeurs de l'entreprise ITESOFT-Yooz avec lesquels je collabore. Ce système manipule les paires de pages consécutives (p_i, p_{i+1}) et extrait des descripteurs structurels et factuels pour déterminer s'il y a continuité ou rupture entre ces pages. En cas de continuité, le système associe les deux pages au même document tandis qu'en cas de rupture, p_i termine le document en cours de traitement et p_{i+1} commence un nouveau document. Ce système achève une précision de 85.6%. Mon travail avait deux contributions principales :
  - d'abord, j'ai renforcé le système par des descripteurs contextuels. En effet, dans le cas où aucun signe structurel (police, taille, layout, listes) ni factuel (pagination, signes de début, signes de fin) n'est identifié, seule la similarité sémantique peut aider à la prise de décision surtout pour les contrats, les conditions générales de vente et tout document contenant beaucoup de

blocs textuels. J'ai testé, par conséquent plusieurs méthodes répandues de calcul de similarité entre les paires de pages. Ces méthodes sont inspirées des modèles de sac de mots, TF*IDF, plongement lexical de mots et doc2vec. Je n'ai pas utilisé des ressources linguistiques comme les lexiques et les dictionnaires puisqu'ils sont très spécifiques et dépendent de la langue utilisée. Les méthodes proposées permettent de convertir le texte de chaque page en vecteur en se basant sur la fréquence des termes présents dans la page. Ces termes sont pondérés selon leur importance dans le document. La similarité entre les pages revient finalement à calculer la distance entre leurs vecteurs. Une étude comparative entre le système à base de règle et les méthodes de similarité textuelle a été publiée dans le workshop ICDAR-WML-2017.

– Deuxièmement, les résultats montrent que dans 92.4% de cas au moins l'un des descripteurs génère une décision correcte. Pour atteindre cette précision, il suffit de mettre tous les descripteurs à plat et sélectionner à chaque fois la bonne décision. J'ai par conséquent entraîné un système de classification qui permet selon les réponses des descripteurs de classifier chaque paire de page en continuité ou rupture. Après classification, la précision du système s'élève à 91.5%. Ce travail a été publié dans la conférence DAS-2018.

Dans le cadre de ce travail, j'ai participé à l'encadrement d'un étudiant en Master 2 informatique, qui s'est poursuivi en thèse. L'encadrement a aboutit à une publication dans le workshop ICDAR-WML-2017.

Mai 2014-Septembre 2014

Poste : Stagiaire scientifique
Unité de recherche : Center for Computational Learning Systems (CCLS)
Lieu : Columbia University (New York - USA)
Superviseurs : Nizar Habash, Owen Rambow
Contributions :

- **Analyse morphologique du dialecte tunisien** : dans le cadre d'un séjour scientifique à l'université de Columbia (New York) en 2014, j'ai étendu un système existant d'analyse et de génération morphologiques de l'arabe et de ses dialectes pour le traitement automatique du tunisien. Ce système relie une représentation morphologique profonde composée d'une racine, d'un schème et d'un ensemble de traits morphologiques à une forme surfacique à travers une série de transformations. Ces transformations sont assurée à l'aide d'un transducteur multi-bandes. Ce système a été utilisé dans mes travaux de thèse afin d'analyser un texte tunisien et de générer un texte en arabe standard.

2011-2014

Poste : Doctorant
Unité de recherche : Laboratoire d'informatique fondamentale de Marseille

(LIF-CNRS)

Lieu : Aix-Marseille Université
Superviseurs : Alexis Nasr, Núria Gala
Contributions :

- **Traitement automatique du dialecte tunisien à partir des ressources de l'arabe standard** : Dans le cadre de mes travaux de thèse, j'ai travaillé sur le traitement automatique des langue peu dotée, qui ne dispose pas d'assez de ressources linguistiques afin de créer des des outils robustes de traitement automatique des langues (TAL). L'objectif est de proposer une approche générique qui permet d'exploiter les ressources existantes d'une langue étymologiquement proche riche en termes de ressources et d'outils de TAL. L'approche proposée consiste à passer par une étape de traduction de la langue peu-dotée vers la langue mieux-dotée puis utiliser un outil TAL existant de la langue cible pour le traitement de la langue source. Dans notre cas, je me suis intéressé à l'étiquetage morphosyntaxique du dialecte tunisien à l'aide d'un étiqueteur de l'arabe moderne standard (MSA). Nous avons traduit le tunisien vers une forme approximative du MSA (pseudo-MSA). Ce dernier n'a pas pour vocation d'être compris par un être humain mais l'utilisation d'un étiqueteur morphosyntaxique destiné au MSA peut fournir des résultats satisfaisants sur le tunisien. De façon plus précise, la traduction que nous proposons repose sur la morphologie et le lexique. Le système proposé relève d'une architecture à transfert. Un mot en tunisien est analysé sous la forme d'une racine, d'un schème et de traits morphologiques. Un lexique bilingue permet alors de convertir la racine et le schème source vers une racine et un schème cible (MSA). La racine et le schème cible, ainsi que les traits morphologiques vont alors permettre de générer un ou plusieurs mots cibles. Un étiqueteur morphosyntaxique entraîné sur des larges ressources existantes en MSA a été ainsi appliqué sur les mots cibles pour assigner les étiquettes morphosyntaxiques adéquates aux mots MSA cibles. Ces étiquettes ont été enfin projetées sur les mots sources tunisiens. Cette méthode a permis d'atteindre une précision de 89% dans l'analyse morphosyntaxique du dialecte tunisien.

Février 2011-Juin 2011

Poste : Stage de fin d'étude de master
Unité de recherche : Laboratoire d'informatique fondamentale de Marseille (LIF-CNRS)

Lieu : Aix-Marseille Université
Superviseurs : Alexis Nasr, Núria Gala
Contributions :

- **Apport de la diacritisation dans l'analyse morphosyntaxique de l'arabe** : La langue arabe se caractérise par l'absence des voyelles courtes (diacritiques) dans la plupart des textes écrits.

En effet, contrairement au français, les voyelles courtes arabes ne sont pas des lettres de l'alphabet, ce sont des signes diacritiques qui se rajoutent aux consonnes (lettres) et qui jouent le même rôle que les voyelles dans les autres langues. La diacritisation en arabe est l'opération qui consiste à attribuer des diacritiques aux lettres des mots non diacrités. Cet exercice est à la fois classique et important dans le traitement automatique de l'arabe. Généralement, les écrits en arabe sont non diacrités et c'est au lecteur de deviner les diacritiques des textes au moment de la lecture. En revanche, les textes religieux et quelques ouvrages scolaires sont entièrement diacrités. D'autres ressources, telles que les textes journalistiques, peuvent être partiellement diacrités. Les diacritiques rajoutés dans ces écrits sont utilisés pour lever des ambiguïtés morphologiques, syntaxiques et parfois sémantiques. Les diacritiques casuels, par exemple, servent à lever l'ambiguïté syntaxique. Ces diacritiques s'associent à la dernière lettre d'un mot à valeur nominale et ils marquent le cas. Ils aident à identifier les fonctions syntaxiques des mots dans une phrase. Les diacritiques affectés aux autres lettres sont appelés lexicaux, ils sont employés pour lever les ambiguïtés morphologiques et sémantiques. L'absence des diacritiques dans un mot provoque des difficultés dans le traitement automatique de l'arabe. C'est-à-dire, qu'un mot non diacrité est plus ambigu qu'un mot partiellement diacrité. Bien que les diacritiques soient destinés à lever les ambiguïtés lors d'un traitement automatique, la majorité des analyseurs morphosyntaxiques de l'arabe n'analysent que des textes non diacrités à cause du manque de ressources arabes diacritées. Par conséquent, si l'entrée est partiellement diacritée, ces analyseurs commencent par élaguer tous les diacritiques, puis ils font l'analyse comme si l'entrée était non diacritée. Les analyseurs morphosyntaxiques de l'arabe n'exploitent donc pas des diacritiques présents dans les textes pour désambiguïser les mots. Dans le cadre de mon travail de stage de fin d'études de master à Aix-Marseille université, j'ai proposé une méthode qui permet de prendre en compte ces diacritiques. Leur prise en compte a naturellement amélioré la diacritisation automatique, la lemmatisation et l'analyse morphologique de l'arabe. Ce travail a été publié dans la conférence nationale RECITAL 2012.

Février 2010-Juillet 2010

Poste : Stage de fin d'études d'ingénieur
Unité de recherche : laboratoire Langage, Langues et Cultures d'Afrique (LLACAN-CNRS)
Lieu : École Nationale de Sciences de l'Informatique (ENSI - Tunisie)
Superviseurs : Fethi Debili
Contributions :

- Étiquetage grammatical de l'arabe standard : dans le cadre de mon projet de fin d'étude à l'école nationale des sciences de l'informatique, j'ai travaillé sur l'étiquetage grammatical de l'arabe. En effet, la

langue arabe, comme toute autre langue, est grammaticalement ambigu. En dehors de tout contexte, chaque mot possède plusieurs étiquettes grammaticales (lemme, partie de discours) potentielles. Cet ensemble d'étiquettes constitue pour chaque mot une liste ambiguë. Mon travail consistait à développer un outil pour attribuer la bonne étiquette grammaticale à chaque mot d'un texte selon le contexte où le mot apparaît. L'outil est basé sur les fréquences de succession des listes ambiguës. Par exemple, si un mot m_1 qui peut potentiellement être verbe ou nom est suivi d'un mot m_2 qui peut être un adjectif ou verbe alors les étiquettes les plus appropriées à ces deux mots sont respectivement nom et adjectif car cette succession est très fréquente. Les fréquence de succession de listes ambiguë sont entraînes à l'aide de corpus préalablement annoté.

Activités d'enseignement

2017-Aujourd'hui **Poste :** Vaccatire
Lieu : La Rochelle Université
Nombre d'heures : 134H
Modules enseignés : Mise en oeuvre des systèmes Big data [CM, TP], Technologies de l'information, Informatique dans l'e-tourisme [CM, TP].

2014-2016 **Poste :** ATER
Lieu : Aix-Marseille Université
Nombre d'heures : 384H
Modules enseignés : Algorithmique, Introduction informatique et programmation, Programmation C II, Outils informatiques et C2i, Programmation Python II, Développement web, Automates et circuits.

2011-2014 **Poste :** Moniteur
Lieu : Aix-Marseille Université
Nombre d'heures : 128H
Modules enseignés : Introduction informatique et programmation, Outils informatiques et C2i, Automates et circuits.

Description des modules enseignés :

- **Mise en œuvre des systèmes big data**: destiné aux étudiants en master (M1) informatique. Mes interventions consistent en 4 heures de cours magistral et 6 heures de TPs à l'université de Niort et La Rochelle université. Mon cours décrit les techniques de bases de plongement lexical de mots (word embedding) et son application au traitement automatique des langues. Depuis 2013, plusieurs modèles de word embedding ont été entraînés à partir de larges collections de Wikipédia, Common Crawl, Twitter, . Ces modèles permettent de convertir les mots d'une langue en vecteurs de nombres réels selon le contexte où chaque mot apparaît. La projection de ces vecteurs dans un espace permet d'encoder le sens des mots et plusieurs relations sémantiques peuvent être déduites à partir des opérations arithmétiques sur les vecteurs (exp: France - Paris + Italie = Rome). La représentation des mots en vecteurs permet aussi de calculer la distance entre eux. Les mots qui partagent le même contexte sont regroupés ensemble dans l'espace et les mots similaires sont généralement représentés par des vecteurs parallèles ou presque d'où la distance cosinus tend vers zéro. J'ai présenté également

aux étudiants l'application de plongements de mots dans plusieurs tâches de traitement automatiques des langues comme l'étiquetage des mots, la reconnaissance des entités nommées. En plus, j'ai démontré aux étudiants comment augmenter ces modèles pour représenter des textes (ensemble de mots) par des vecteurs ce qui permet de calculer la similarité textuelle ou de faire la classification textuelle (analyse de sentiments...). En TP, les étudiants ont entraîné leurs propres modèles à l'aides de petits jeux de données et ont aussi utilisé des modèles pré-entraînés existants afin de reproduire les relations sémantiques vues dans le cours. Enfin les étudiants ont appliqué des fonctions pré-définies pour des tâches de traitement automatique des langues comme la classification de textes, l'étiquetage en parties de discours, la reconnaissance d'entités nommées et le résumé automatique.

- **Informatique dans l'e-tourisme**: destiné aux étudiants de toutes les licences et souhaiteraient, (1) soit s'orienter vers une formation tourisme à l'issue de leur licence, (2) soit entrer dans la vie professionnelle à l'issue de leur Bac+3. Ce cours permet aux étudiants de donner une coloration à leur formation et de découvrir le monde professionnel du tourisme. Ainsi, des étudiants des différentes licences pourront découvrir les métiers et compétences nécessaires pour une insertion professionnelle réussie dans ce secteur d'activité. Mes interventions analysent la reconnaissance d'entités nommées (exp. noms d'hôtels), la détection d'opinions et leur application dans l'e-tourisme.

- **Technologies de l'information**: destiné aux étudiants en L2 gestion. L'objectif est de concevoir des pages web statistiques et dynamiques à l'aide des langages HTML, CSS, JavaScript, PHP et encore en utilisant des CMS disponibles sur Internet.

- **Algorithmique**: destiné aux étudiants en L2 informatique, ou en option aux étudiants en L2 mathématiques. Le cours introduit les notions fondamentales de preuves et de complexité des algorithmes ainsi que les structures de données linéaires (tableaux, listes, piles, files et tables de hachage) et les graphes. Ce cours explique également la programmation dynamique qui permet de résoudre des problèmes d'optimisation.

- **Introduction informatique et programmation**: destiné aux étudiants en L1 informatique, introduit les notions de base nécessaires pour écrire des programmes simples en langage C : instructions, variables et types simples, structures conditionnelles, boucles itératives, tableaux, fonctions et passage de paramètres, manipulation des fichiers.

- **Programmation C II**: destiné aux étudiants en L1 (semestre 2) et en L2 (semestre 1) informatique. Ce cours complète l'introduction informatique et programmation. Il propose un apprentissage avancée en programmation C avec l'introduction des types complexes, des pointeurs et de la récursivité.

- **Outils informatiques et C2i**: destiné aux étudiants en L1 informatique. L'objectif de ce cours est de se familiariser avec des outils informatiques de base (LaTeX, Excel et Unix shell).

- **Programmation Python II**: destiné aux étudiants des Classes Préparatoires aux Grandes Écoles (CPGE) à Aix-Marseille Université. Ce cours permet l'étude de la programmation fonctionnelle, procédurale et orientée objet. Il permet d'explorer également des fonctions pré-définies dans des librairies Python à l'instar de NumPy et Matplotlib.

- **Développement web**: destiné aux étudiants en L1 informatique. L'objectif de ce cours est d'apprendre à créer des pages web à l'aide de HTML, CSS, JAVASCRIPT et JQUERY.

- **Automates et circuits**: destiné aux étudiants en L1 informatique. Ce cours introduit les systèmes de numération et le codage d'informations. Il propose des rappels sur les ensembles, le dénombrement, les fonctions et les relations. Enfin, il permet d'étudier les circuits à mémoire et bascules et les automates.

Récapitulation des activités d'enseignement :

Année	Intitulé	Public	Spécialité	Cours	TD	TP
2020-2021	Mise en œuvre des systèmes big data	Master 1	Informatique	4	–	6
	Informatique dans l'e-tourisme I	toutes les licences		6	6	–
	Technologies de l'information	Licence 1	Gestion	–	28	–
2019-2020	Technologies de l'information	Licence 1	Gestion	–	28	–
2018-2019	Technologies de l'information	Licence 1	Gestion	–	28	–
2017-2018	Technologies de l'information	Licence 1	Gestion	–	14	–
2016-2017	Technologies de l'information	Licence 1	Gestion	–	14	–
2015-2016	Algorithmique	Licence 2	Mathématiques	–	24	20
	Introduction informatique et programmation	Licence 1	Informatique	20	24	21
	Outils informatiques et C2I	Licence 1	Informatique	–	24	–
	Programmation Python II	Licence 1	Informatique	–	24	–
	Automates et circuits	Licence 1	Informatique	–	24	18
2014-2015	Algorithmique	Licence 2	Informatique	–	24	20
	Outils informatiques et C2I	Licence 1	Informatique	–	48	–
	Introduction informatique et programmation	Licence 1	Informatique	–	–	63
	Programmation C II	Licence 1	Informatique	–	18	21
	Développement web I	Licence 1	Informatique	–	15	–
2013-2014	Introduction informatique et programmation	Licence 1	Informatique	–	–	42
	Outils informatiques et C2I	Licence 1	Informatique	–	24	–
2012-2013	Introduction informatique et programmation	Licence 1	Informatique	–	–	21
	Outils informatiques et C2I	Licence 1	Informatique	–	24	–
	Automates et circuits	Licence 1	Informatique	–	24	–

Dans le cadre de mes activités d'enseignement, j'ai pu assurer des diverses modules pour plusieurs niveaux et disciplines. Pour mes prochaines interventions, je souhaite enseigner d'autres modules pour varier mon expérience. Je souhaite aussi apporter une attention particulière aux modules liés à la linguistique computationnelle, à l'extraction d'information et à l'apprentissage automatique, thèmes sur lesquels se basent mes activités de recherche, afin d'appliquer mon expertise de recherche dans l'enseignement. Je tiens également à notifier que dans tous mes enseignements, j'étais responsable dans la correction des partiels et/ou des examens.

Publications

1. **Conférences internationales**
   - SIGIR 2021 (Core A*).
     A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. **Ahmed Hmadi**, Elvys Linhares Pontes, Emanuela Boros, Tuyet Hai Nguyen Thi, Hackl Günter, Jose G. Moreno et Antoine Doucet. The 44^{th} International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021).
   - ICDAR 2020 (Core A).
     Information extraction from invoices. **Ahmed Hmadi**, Elodie Carel, Aurélie Joseph, Mickaël Coustaty et Antoine Doucet. The 16^{th} International Conference on Document Analysis and Recognition.

- TPDL 2020 (Core B).
  Assessing and Minimising the Impact of OCR Quality on Named Entity Recognition. **Ahmed Hmadi**, Axel-Jean Caurant, Nicolas Sidère, Mickaël Coustaty et Antoine Doucet. Digital Libraries for Open Knowledge - 24^{th} International Conference on Theory and Practice of Digital Libraries. Springer, 2020, pp. 87–101. DOI: 10.1007/978-3-030-54956-5_7. https://doi.org/10.1007/978-3-030-54956-5_7 [SLIDES]
- CONLL 2020 (Core A).
  Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. Emanuela Boros, **Ahmed Hamdi**, Elvys Linhares Pontes, Luis Adrian Cabrera-Diego, Jose G. Moreno, Nicolas Sidère et Antoine Doucet. Proceedings of the 24^{th} Conference on Computational Natural Language Learning, pp. 431–441. https://www.aclweb.org/anthology/2020.conll-1.35
- ICADL 2020 (Core A).
  When to Use OCR Post-correction for Named Entity Recognition? Vinh-Nam Huynh, **Ahmed Hamdi** et Antoine Doucet. Digital Libraries at Times of Massive Societal Transition - 22^{nd} International Conference on Asia-Pacific Digital Libraries. Lecture Notes in Computer Science 12504 (2020), pp, 33-42. DOI: 10.1007/978-3-030-64452-9_3. https://doi.org/10.1007/978-3-030-64452-9_3. [SLIDES]
- ICADL 2020 (Core A).
  Entity Linking for Historical Documents: Challenges and Solutions. Elvys Linhares Pontes, Luis Adrian Cabrera-Diego, Jose G. Moreno, Emanuela Boros, **Ahmed Hamdi**, Nicolas Sidère, Mickaël Coustaty et Antoine Doucet. Digital Libraries at Times of Massive Societal Transition - 22^{nd} International Conference on Asia-Pacific Digital Libraries. Lecture Notes in Computer Science 12504 (2020), pp, 215-231. DOI: 10.1007/978-3-030-64452-9_19. https://doi.org/10.1007/978-3-030-64452-9_19.
- JCDL 2019 (Croe A*)
  An Analysis of the Performance of Named Entity Recognition over OCRed Documents. Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty et Antoine Doucet. 19^{th} ACM/IEEE Joint Conference on Digital Libraries, pp, 333-334. DOI: 10.1109/JCDL.2019.00057. https://doi.org/10.1109/JCDL.2019.00057.
- ICADL 2019 (Core A).
  Impact of OCR Quality on Named Entity Linking. Elvys Linhares Pontes, **Ahmed Hamdi**, Nicolas Sidere et Antoine Doucet. Digital Libraries at the Crossroads of Digital Information for the Future - 21^{st} International Conference on Asia-Pacific Digital Libraries.Lecture Notes in Computer Science 11853 (2019) pp, 102-11. DOI: 10.1007/978-3-030-34058-2_11. https://doi.org/10.1007/978-3-030-34058-2_11,
- DAS 2018 (Core B).
  Feature Selection for Document Flow Segmentation. **Ahmed Hamdi**, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain D'Andecy, Antoine Doucet et Jean-Marc Ogier. 13^{th} IAPR International Workshop on Document Analysis Systems. IEEE Computer Society, pp, 245–250. DOI:10.1109/DAS.2018.66. https://doi.org/10.1109/DAS.2018.66.

- MT SUMMIT 2013 (Core B).
  The Effects of Factorizing Root and Pattern Mapping in Bidirectional Tunisian - Standard Arabic Machine Translation. **Ahmed Hamdi**, Rahma Boujelbane, Nizar Habash et Alexis Nasr. MT Summit 2013. `https://hal.archives-ouvertes.fr/hal-00908761`.

2. **Workshops internationaux**
   - CEUR@CLEF 2020
     Robust Named Entity Recognition and Linking on Historical Multilingual Documents. Emanuela Boros, Elvys Linhares Pontes, Luis Adrian Cabrera-Diego, **Ahmed Hamdi**, Jose G. Moreno, Nicolas Sidère et Antoine Doucet. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, 2020. `http://ceur-ws.org/Vol-2696/paper_171.pdf`.
   - WML@ICDAR 2017
     Machine Learning vs Deterministic Rule-Based System for Document Stream Segmentation. **Ahmed Hamdi**, Joris Voerman, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain D'Andecy et Jean-Marc Ogier. First Workshop of Machine Learning, 14^{th} IAPR International Conference on Document Analysis and Recognition, WML@ICDAR 2017. IEEE, pp, 77-82. DOI: 10.1109/ICDAR.2017.332 `https://doi.org/10.1109/ICDAR.2017.332`.
   - WANLP@ACL 2015
     POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools. **Ahmed Hamdi**, Alexis Nasr, Nizar Habash et Nuria Gala. Proceedings of the 2^{nd} Workshop on Arabic Natural Language Processing Processing, ANLP@ACL 2015. Association for Computational Linguistics. DOI: 10.18653/v1/W15-3207. `https://doi.org/10.18653/v1/W15-3207`.
   - VarDial@COLING 2014
     Automatically building a Tunisian Lexicon for Deverbal Nouns. **Ahmed Hamdi**, Nuria Gala et Alexis Nasr. Proceedings of the 1^{st} Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, VarDial@COLING 2014. Association for Computational Linguistics and Dublin City University, pp, 95-102. DOI: 10.3115/v1/W14-5311. `https://doi.org/10.3115/v1/W14-5311`.

3. **Livrables de projets**
   - *Named Entity Recognition and Linking.* **Ahmed Hamdi**, Elvys Linhares Pontes et Antoine Doucet. NewsEye 2020.
   - *Stance detection.* **Ahmed Hamdi**, Thi Tuyet Haï Nguyen et Antoine Doucet. NewsEye 2020.

4. **Rapport de thèse**
   - Ahmed Hamdi, 2015. *Traitement automatique du dialecte tunisien à l'aide d'outils et de ressources de l'arabe standard : application à l'étiquetage morphosyntaxique.* Aix-Marseille Université. `https://www.theses.fr/en/2015AIXM4089`.

5. **Conférences nationales**
   - CORIA 2021
     Atténuer les erreurs de numérisation dans la reconnaissance d'entités nommées pour les documents historiques. Emanuela Boros, **Ahmed Hamdi**, Elvys Linhares Pontes, Luis Adrian Cabrera-Diego, Jose G. Moreno, Nicolas Sidère et Antoine Doucet. Conférence francophone en Recherche d'Information et Application.
   - TALN 2013
     Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. **Ahmed Hamdi**, Rahma Boujelbane, Nizar Habash et Alexis Nasr. Traitement Automatique des Langues Naturelles, TALN 2013. The Association for Computer Linguistics, pp, 395-406. `https://www.aclweb.org/anthology/F13-1029/`.
   - RECITAL 2012
     Apport de la diacritisation de l'analyse morphosyntaxique de l'arabe. **Ahmed Hamdi**. Proceedings of the Joint Conference JEP-TALN-RECITAL 2012. ATALA/ AFCP pp, 247-254. `https://www.aclweb.org/anthology/F12-3019/`.

Responsabilités collectives et administratives

- Participation à l'organisation de la conférence internationale CICLING (*International Conference on Intelligent Text Processing and Computational Linguistics*) 2019 à La Rochelle. `https://www.cicling.org/2019/#Conference_committees`
- Participation à l'organisation de la Journée des Ingénieurs, Doctorants, ATER et Post-docs (JIDAP) en 2018 et 2019. La JIDAP est un événement annuel au laboratoire L3i qui a pour objectif de favoriser les échanges entre les Ingénieurs, Doctorants, ATER et PostDoc qui permettrait à chacun d'avoir une idée sur les thèmes de recherche de ses collègues et de découvrir plein de choses utiles sur le laboratoire particulièrement aux nouveaux venants.
- Participation aux fêtes de la science 2018, 2019 et 2020: la fête de la science est un événement national proposé par le Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation où les chercheurs scientifiques de La Rochelle Université présentent leurs travaux au public (y compris les écoliers, collégiens et lycéens) afin de satisfaire toutes les curiosités. Dans ce cadre, j'ai pu présenter mes travaux de recherche dans la segmentation de flux documentaires en 2018 et l'extraction d'informations à partir de documents administratifs en 2019. Enfin, en 2020, j'ai présenté mes travaux sur la reconnaissance et la désambiguïsation des entités nommées à partir de la presse ancienne. `https://www.univ-larochelle.fr/wp-content/uploads/pdf/FDLS-2020-PROGRAMME-PUBLIC-20200923.pdf`
- Participation aux journées portes ouvertes 2019 et 2020: les journées portes ouvertes sont des évènements organisés par l'université afin de présenter aux futur(e)s étudiant(e)s les différents travaux dans les unités de recherche rattaché à La Rochelle université. Ces journées font aussi l'occasion pour visiter les locaux et échanger avec les enseignants sur les formations dispensées et les services proposés. [SLIDES JPO 2020]

Récompenses et prix

- CLEF-HIPE 2020: Named Entity Recognition (NER) and Entity Linking (EL) on Historical Newspapers NER,
  - NER: **l'équipe L3i est première sur tous les classements**
  - EL: **l'équipe L3i est première sur 50/52 classements et deuxième sur 2 classements**
    `https://github.com/impresso/CLEF-HIPE-2020/blob/master/evaluation-results/ranking_summary_final.md`
- ICADL 2019: "Impact of OCR quality on named entity linking", **best paper**
- Hackathon "La Rochelle ville connectée" 2017, **prix du Hackathon**
  `https://opendata.larochelle.fr/1er-hackathon-une-grande-reussite/`