

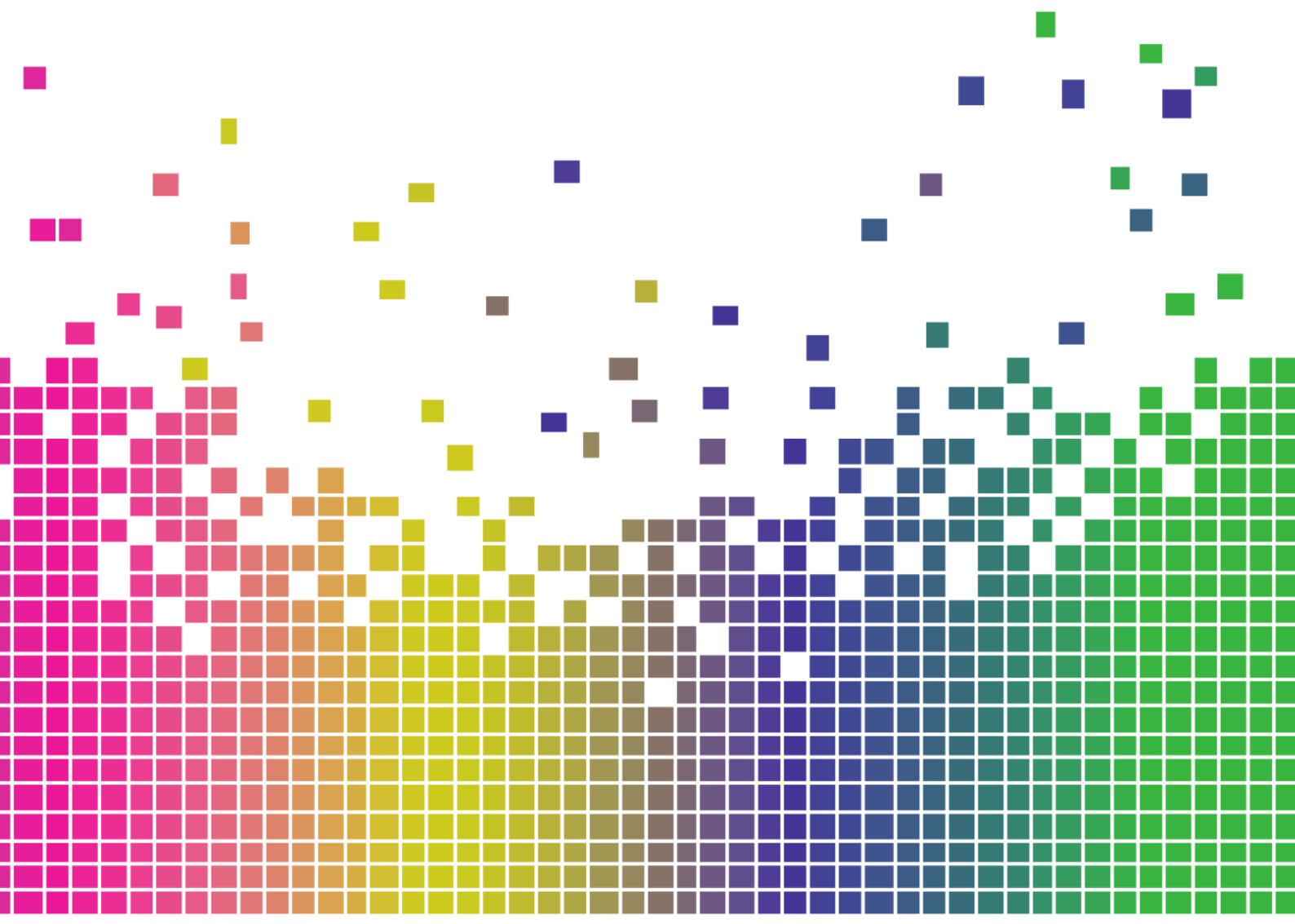


Rennes
14 - 18 mai

Actes de la conférence TALN 2018

Volume 1 : Articles longs, articles courts de TALN

Pascale Sébillot, IRISA, INSA Rennes
Vincent Claveau, CNRS, IRISA, Univ. Rennes



Préface

Mots des présidents des comités de programme

Pour la première fois, l'ARIA (Association francophone de Recherche d'Information et Applications) et l'ATALA (Association pour le Traitement Automatique des Langues) ont organisé conjointement leur principale conférence annuelle afin de réunir en un seul lieu les deux communautés de la recherche d'information (RI) et du traitement automatique des langues (TAL). Organisée par l'IRISA (UMR 6074) et le Centre Inria Bretagne-Atlantique, cette édition s'est déroulée du 14 au 18 mai 2018 à Rennes. Elle a donc regroupé :

- la 15^{ème} Conférence en Recherche d'Information et Applications (CORIA) ;
- la 25^{ème} conférence sur le Traitement Automatique des Langues Naturelles (TALN) ;
- une rencontre jeunes chercheurs (RJC) commune aux deux communautés correspondant à la 13^{ème} édition de la Rencontre des Jeunes Chercheurs en Recherche d'Information (RJCRI) et à la 20^{ème} édition des Rencontre Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) ;
- le salon de l'innovation en technologies du langage et de l'information.

Des ateliers et tutoriels, un hackathon ainsi qu'un salon de l'innovation à destination des industriels ont aussi enrichi ce programme (voir plus bas).

Les actes de CORIA ne sont pas présents dans ce volume mais sont accessibles à <http://www.asso-aria.org/>. Cette année, il y avait un seul format de soumission à CORIA mais deux formats de présentation pour les articles acceptés. Nous avons acceptés quinze articles pour une présentation longue et quatre articles pour des présentations courtes. Le taux de sélection pour les articles en présentation longue est de cinquante pourcent. Vingt-trois villes différentes sont représentées dans les dix-neuf papiers acceptés; beaucoup de travaux sont issus de collaborations, dont certaines internationales. Six papiers acceptés ont un auteur d'une organisation située à Toulouse, cinq d'une organisation parisienne et trois d'une organisation grenobloise. Au niveau international, nous pouvons noter des contributions acceptées provenant du Canada, de Russie et de Tunisie. Nous pouvons également noter des soumissions provenant du Cameroun et de Madagascar. La majorité des articles proviennent de laboratoires de recherche académiques. Les thèmes abordés à la fois dans les soumissions et dans les papiers acceptés sont variés tant au niveau des questions de recherche abordées que des méthodes proposées pour les résoudre et des collections utilisées pour valider ou évaluer les propositions.

Cette année, vingt deux articles ont été soumis à RJC. Après avoir été chacun évalué par trois membres du comité de programme, quatre articles ont été retenus pour une présentation orale (soit un taux de sélection pour présentation orale de 18 %), et neuf autres ont été retenus pour une présentation sous forme de poster (taux de sélection global de 59 %). Nous avons ainsi pu donner l'opportunité à treize jeunes chercheuses et chercheurs, en grande majorité en début de thèse, de présenter leurs travaux à la communauté.

Cette année, TALN inaugurait de nouvelles modalités de soumissions : un appel unique et un seul format de soumission en article court pouvant être étendu en article long sur proposition du comité de programme. Parmi les soixante douze articles soumis suite à cet appel, le comité de programme a proposé à quatorze d'entre eux un passage en format long (soit un taux de sélection de 19,5 %) et en a retenu quarante deux autres en articles courts. Pour effectuer cette sélection, le comité de programme s'est appuyé sur trois à quatre relectures effectuées par des membres du comité de lecture (liste donnée ci-après), synthétisées et portées lors de la réunion du comité de programme par les



Figure 1: Nuage de termes extraits des actes de TALN.

responsables de domaine. L'ensemble de ce processus s'est déroulé comme les années précédentes en double aveugle. Les nombre de soumissions et le taux de sélection placent ainsi cette édition dans les pas de celles des années précédentes, suivant le double objectif d'avoir une conférence conservant d'une part une sélectivité forte, garante de la qualité des interventions orales, et se voulant, d'autre part, également un lieu de rencontre le plus ouvert possible à l'expression de l'ensemble de la communauté, au travers des articles courts.

Les thématiques abordées dans les articles retenus dans ces conférences sont variées. Sans surprise, les tendances de fond que constituent l'apprentissage profond et les plongements lexicaux occupent une part importante des contributions, mais pour autant d'autres approches et de nombreux domaines sont explorés. Les sessions ont ainsi porté sur les domaines d'application particuliers (domaines de spécialité, langues peu dotées), des niveaux d'analyses linguistiques (morphologie, syntaxe, lexique) ou des tâches spécifiques (résumé automatique, OCR, multimédia, fouille d'opinion). La figure 1 présente un nuage de termes extraits de ces actes¹.

En complément de ce programme, nous avons eu l'honneur d'accueillir deux oratrices invitées reconnues internationalement : Dina Demner (NUH, US National Library of Medicine) qui a présenté des avancées récentes en traitement automatique du langage biomédical, et Claudia Hauff (TU Delft) qui a effectué un exposé sur l'apprentissage humain en recherche d'information. Il convient également de citer le salon de l'innovation, qui, avec ses tables rondes, démonstrations, stands d'industriels du secteur ou de projets de recherche, permet aux industriels et aux chercheurs en TAL et RI, ainsi qu'aux entreprises en technologie de l'information et plus généralement du numérique, de se rencontrer

¹Outils : TermEx, disponible sur <https://algo.inria.fr>, et <https://www.wordclouds.com>.

et d'échanger autour des idées de développements actuels et futurs du domaine, de promouvoir les enjeux et applications du secteur, ainsi que de renforcer la visibilité et l'image des entreprises, organisations, institutions et projets de recherche auprès de partenaires et clients potentiels. Enfin, les conférences ont été précédées de deux journées d'ateliers et tutoriels se focalisant sur certaines thématiques plus précises du TAL et de la RI, portant sur la recherche d'information sémantique (atelier RISE), la fouille de texte (défi DeFT, cette année sur l'analyse de sentiment, dont les actes sont proposées dans le second volume du présent ouvrage), l'analyse des données de la recherche (atelier VaDOR), l'infrastructure de fouille de texte européenne OpenMinTed (tutoriel), le hackathon sur les fausses nouvelles ou infox (*fake news*), l'analyse des réseaux sociaux (atelier ALIAS, soutenu par le GdR CNRS MaDICS), et le data-journalisme (atelier CAJOLE, soutenu par le GdR CNRS MaDICS).

P. Cellier (RJC), A.-L. Ligozat (RJC), J. Mothe (CORIA), P. Sébillot (TALN), V. Claveau (TALN)

Mots des Présidents de l'ATALA et de l'ARIA

Cette année, les associations ARIA et ATALA ont souhaité organiser conjointement leur conférence à Rennes. L'objectif était de permettre aux chercheurs des deux communautés de se retrouver en un même lieu et un même temps autour de thématiques qu'ils partagent. En effet, le domaine de la Recherche d'Information ayant pour objectif d'identifier les informations les plus appropriées par rapport au besoin d'un usager, il repose sur différentes stratégies parmi lesquelles les modèles de langue trouvent une place spécifique. De même, dans le domaine du Traitement Automatique des Langues la transition du papier au support électronique nécessite des fonctionnalités se rapprochant de plus en plus des compétences humaines, phénomène amplifié par le retour sur le devant de la scène scientifique de l'Intelligence Artificielle, accompagné d'une demande croissante pour des agents informatiques donnant l'illusion de l'autonomie linguistique. Cette coïncidence n'est pas surprenante car la RI et le TAL partagent dès le début de l'informatique, une histoire commune avec l'IA, n'oublions pas en effet que les mesures d'évaluation emblématiques de la RI que sont la précision et le rappel ont été élaborées en 1960, lors des expériences du College of Aeronautics de Cranfield (UK) et qu'à la même époque, la communauté TAL se constituait autour de la traduction automatique, avec la naissance de l'ATALA à Paris, en 1959. Pour ce qui concerne l'IA, beaucoup considèrent l'atelier qui s'est tenu au Dartmouth College (USA) en 1956 comme marquant la naissance du domaine. C'est aussi pendant les années 60 que l'on a vu apparaître les premières implémentations d'algorithmes neuromimétiques pour l'apprentissage automatique. Nos deux communautés partageant des débuts contemporains et étant unies comme par le passé autour de problématiques communes, nous avons donc fait le choix, cette année, de favoriser les échanges et les présentations communes au travers de l'organisation de ces conférences conjointes.

P. Paroubek (ATALA) & M. Chevalier (ARIA)

Comité d'organisation de CORIA-TALN-RJC

Coordinateur :

Vincent Claveau, CNRS, IRISA, Univ. Rennes

Webmestres :

Clément Dalloux, CNRS, IRISA, Univ. Rennes

Cédric Maigrot, IRISA, Univ Rennes

Resp. démonstrations :

Anne-Lyse Minard, CNRS, IRISA, Univ. Rennes

Resp. ateliers :

Annie Forêt, IRISA, Univ. Rennes

Resp. salon de l'innovation :

Géraldine Damnati, Orange, Lannion

Aleksandra Gerraz, Orange, Lannion

Resp. sponsoring :

Gwénolé Lecorvé, IRISA, ENSSAT, Univ. Rennes

Infographiste :

Agnès Cottais, IRISA, Rennes

Support administratif :

Élisabeth Lebret, Inria, Rennes

Aurélie Patier, IRISA, Rennes

Membres du comité d'organisation :

Cheikh Brahim El Vaigh, Inria, Rennes

Peggy Cellier, IRISA, INSA Rennes

Guillaume Gravier, IRISA, CNRS, Rennes

Pierre-François Marteau, IRISA, Univ. Bretagne Sud, Vannes

Nicolas Béchet, IRISA, IUT de Vannes

Pascale Sébillot, IRISA, INSA Rennes

Mikaïl Demirdelen, IRISA, INSA Rennes

Ainsi que les équipes techniques et administratives du centre Inria Rennes Bretagne Atlantique.

Comité de programme TALN

Présidents du comité de programme :

- Pascale Sébillot, IRISA, INSA Rennes
- Vincent Claveau, CNRS, IRISA, Univ. Rennes

Responsables de domaine :

Maxime Amblard, LORIA, Université de Lorraine

Delphine Bernhard, LiLPa, Université de Strasbourg

Philippe Blache, LPL, CNRS

Nathalie Camelin, LIUM, Université du Maine

Iris Eshkol-Taravella, MoDyCo, Université Paris Nanterre

Cécile Fabre, ERSS, Université Toulouse 2

Benoît Favre, LIF, Aix Marseille Université

Olivier Ferret, CEA LIST

Thierry Hamon, LIMSI, Université Paris Nord

Philippe Langlais, RALI/DIRO, Univ. de Montréal

Emmanuel Morin, LS2N, Université de Nantes

Philippe Muller, IRIT, Université Paul Sabatier

Aurélie Névéol, LIMSI, CNRS

Didier Schwab, LIG, Université Grenoble Alpes

Xavier Tannier, LIMICS, Université Pierre et Marie Curie

Comité de lecture :

Stergos Afantenos, IRIT, Université Paul Sabatier

Salah Ait-Mokhtar, NaverLabs

Alexandre Allauzen, LIMSI, Université Paris-Sud

Jean-Yves Antoine, LI, Université Tours

Frédéric Béchet, LIF, Aix Marseille Université

Laurent Besacier, LIG, Université Grenoble Alpes

Romaric Besançon, CEA LIST

Pierrette Bouillon, ETI/TIM/ISSCO, Université de Genève

Chloé Braud, LORIA, CNRS

Marie Candito, LLF, Université Paris Diderot

Thierry Charnois, LIPN, Université Paris 13

Chloé Clavel, Télécom Paris

Guillaume Cleuziou, LIFO, Université Orléans

Mathieu Constant, ATILF, Université Lorraine

Benoît Crabbé, LLf, Université Paris Diderot

Béatrice Daille, LS2N, Université Nantes

Laurence Danlos, LLF, Université Paris Diderot

Marco Dinarelli, LaTTiCe, CNRS

Patrick Drouin, RALI-DIRO, Université de Mon-

tréal

Thomas François, CENTAL, Université catholique de Louvain

Nathalie Friburger, LI, Université Tours

Claire Gardent, LORIA, CNRS

Éric Gaussier, LIG, Université Grenoble Alpes

Natalia Grabar, STL, CNRS

Camille Guinaudeau, LIMSI, Université Paris-Sud

Nabil Hathout, ERSS, Université de Toulouse

Nicolas Hernandez, LS2N, Université de Nantes

Stéphane Huet, LIA, Université d'Avignon et des pays de Vaucluse

Sylvain Kahane, Modyco, Université Paris Ouest - Nanterre

Olivier Kraif, LIDILEM, Université Grenoble Alpes

Mathieu Lafourcade, LIRMM, Université de Montpellier

Guy Lapalme, RALI-DIRO, Université de Montréal

Francois Lareau, OLST, Université de Montréal
Jean-Marc Lecarpentier, Greyc, Université Caen
Basse-Normandie
Gwénolé Lecorvé, IRISA, ENSSAT, Université
de Rennes
Joseph Le Roux, LIPN, Université Paris 13
Anaïs Lefeuvre-Halftermeyer, LIFO, Université
d'Orléans
Sébastien Le Maguer,
Anne-Laure Ligozat, LIMSI, ENSIIE
Denis Maurel, LI, Université Tours
Anne-Lyse Minard, CNRS, IRISA, Université de
Rennes
Richard Moot, LIRMM, CNRS
Véronique Moriceau, LIMSI, Université Paris-
Sud
Adeline Nazarenko, LIPN, Université Paris Nord
Jian-Yun Nie, RALI-DIRO, Université de Mon-
tréal
Yannick Parmentier, LORIA, Université Lorraine
Sylvain Pogodalla, LORIA, INRIA
Thierry Poibeau, LaTTiCe, CNRS

Andrei Popescu-Belis, HEIG VD - School of
Business and Engineering Vaud
Jean-Philippe Prost, LIRMM, Université Mont-
pellier 2
Solen Quiniou, LS2N, Université Nantes
Christian Raymond, IRISA, INSA Rennes
Christian Retoré, LIRMM, Université de Mont-
pellier
Mathieu Roche, CIRAD
Sophie Rosset, LIMSI, CNRS
Michel Simard, National Research Council
Canada (NRC)
Ludovic Tanguy, ERSS, Université Toulouse 2
Isabelle Tellier, Lattice, Université Paris 3
Juan-Manuel Torres-Moreno, LIA, Université
d'Avignon et des pays de Vaucluse
Julien Velcin, ERIC, Université Lyon 2
Guillaume Wisniewski, LIMSI, Université Paris-
Sud
François Yvon, LIMSI, Université Paris-Sud
Pierre Zweigenbaum, LIMSI, CNRS

Table des matières

Préface	iii
Articles longs	
Étude de la lisibilité des documents de santé avec des méthodes d’oculométrie. <i>Natalia Grabar, Emmanuel Farce et Laurent Sparrow</i>	3
Alignement de termes de longueur variable en corpus comparables spécialisés. <i>Jingshu Liu, Emmanuel Morin et Sebastián Peña Saldarriaga</i>	19
Étude de la reproductibilité des word embeddings : repérage des zones stables et instables dans le lexique. <i>Bénédicte Pierrejean et Ludovic Tanguy</i>	33
Modeling infant segmentation of two morphologically diverse languages. <i>Georgia Rengina Loukatou, Sabine Stoll, Damian Blasi et Alejandrina Cristia</i>	47
Évaluation morphologique pour la traduction automatique: adaptation au français. <i>Franck Burlot et François Yvon</i>	61
Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux. <i>Pierre Magistry, Anne-Laure Ligozat et Sophie Rosset</i>	75
Modélisation des processus d’acquisition syntaxique par jeux de langage entre agents artificiels. <i>Marie Marcia et Isabelle Tellier</i>	87
MOTS : un outil modulaire pour le résumé automatique. <i>Valentin Nyzam, Christophe Rodrigues et Aurélien Bossard</i>	101
Ordonnancement de réponses dans les systèmes de dialogue basé sur une similarité contexte/réponse. <i>Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin et Emmanuel Morin</i>	115
Intégration de contexte global par amorçage pour la détection d’événements. <i>Dorian Kodolja, Romaric Besançon et Olivier Ferret</i>	129
Construction conjointe d’un corpus et d’un classifieur pour les registres de langue en français. <i>Gwénolé Lecorvé, Hugo Ayats, Fournier Benoît, Jade Mekki, Jonathan Chevelu, Delphine Battistelli et Nicolas Béchet</i>	143

Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. <i>Loïc Vial, Benjamin Lecouteux et Didier Schwab</i>	157
Correction automatique d'attachements prépositionnels par utilisation de traits visuels. <i>Sébastien Delecraz, Leonor Becerra-Bonache, Benoît Favre, Alexis Nasr et Frédéric Bechet</i>	171
Décodeur neuronal pour la transcription de documents manuscrits anciens. <i>Adeline Granet, Emmanuel Morin, Harold Mouchère, Solen Quiniou et Christian Viard-Gaudin</i>	183

Articles courts

A prototype dependency treebank for Breton. <i>Francis M. Tyers et Vinit Ravishankar</i>	197
Détection automatique de phrases en domaine de spécialité en français. <i>Arthur Boyer et Aurélie Névéol</i>	205
Des représentations continues de mots pour l'analyse d'opinions en arabe: une étude qualitative. <i>Amira Barhoumi, Nathalie Camelin et Yannick Estève</i>	215
Evaluation automatique de la satisfaction client à partir de conversations de type «chat» par réseaux de neurones récurrents avec mécanisme d'attention. <i>Jeremy Auguste, Delphine Charlet, Géraldine Damnati, Benoit Favre et Frederic Bechet</i>	225
Détection d'erreurs dans des transcriptions OCR de documents historiques par réseaux de neurones récurrents multi-niveau. <i>Thibault Magallon, Frederic Bechet et Benoit Favre</i>	233
Le benchmarking de la reconnaissance d'entités nommées pour le français <i>Jungyeul Park</i>	241
Une note sur l'analyse du constituant pour le français <i>Jungyeul Park</i>	251
Interface syntaxe-sémantique au moyen d'une grammaire d'arbres adjoints pour l'étiquetage sémantique de l'arabe <i>Cherifa Ben Khelil, Chiraz Ben Othmane Zribi, Denys Duchier et Yannick Parmentier</i>	261
FinSentiA: Sentiment Analysis in English Financial Microblogs <i>Thomas Gaillat, Annanda Sousa, Manel Zarrouk et Brian Davis</i>	271
L'optimisation du plongement de mots pour le français : une application de la classification des phrases <i>Jungyeul Park</i>	281
Word2Vec vs LSA pour la détection des erreurs orthographiques produisant un dérèglement sémantique an arabe <i>Chiraz Ben Othmane Zribi</i>	293
Analyse de sentiments à base d'aspects par combinaison de réseaux profonds : application à des avis en Français <i>Nihel Kooli et Erwan Pigneul</i>	303
Predicting the Semantic Textual Similarity with Siamese CNN and LSTM <i>Elyvs Linhares Pontes, Stéphane Huet, Andréa Carneiro Linhares et Juan-Manuel Torres-Moreno</i>	311

L'évaluation des représentations vectorielles de mots en utilisant WordNet <i>Nourredine Aliane, Jean-Jacques Mariage et Gilles Bernard</i>	321
Traduction automatique de corpus en anglais annotés en sens pour la désambiguïisation lexicale d'une langue moins bien dotée, l'exemple de l'arabe <i>Marwa Hadj Salah, Loïc Vial, Hervé Blanchon, Mounir Zrigui et Didier Schwab</i>	329
Détection de mésusages de médicaments dans les réseaux sociaux <i>Elise Bigeard, Natalia Grabar et Frantz Thiessard</i>	337
Utilisation de Représentations Distribuées de Relations pour la Désambiguïisation d'Entités Nommées <i>Nicolas Wagner, Romaric Besançon et Olivier Ferret</i>	347
Traduction automatique du japonais vers le français Bilan et perspectives <i>Raoul Blin</i>	357
Des pseudo-sens pour améliorer l'extraction de synonymes à partir de plongements lexicaux <i>Olivier Ferret</i>	365
Annotation automatique des types de discours dans des livres audio en vue d'une oralisation par un système de synthèse <i>Aghilas Sini, Elisabeth Delais-Roussarie et Damien Lolive</i>	375
Impact du prétraitement sur l'Analyse de Sentiment du Dialecte Tunisien <i>Chedi Bechikh Ali, Hala Mulki et Hatem Haddad</i>	383
Detecting context-dependent sentences in parallel corpora <i>Rachel Bawden, Thomas Lavergne et Sophie Rosset</i>	393
Predicting failure of a mediated conversation in the context of asymmetric role dialogues <i>Romain Carbou, Delphine Charlet, Géraldine Damnati, Frédéric Landragin and Jean Léon Bouraoui</i>	401
Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien <i>Clément Dalloux, Vincent Claveau, Natalia Grabar et Claudia Moro</i>	409
Le corpus PASTEL pour le traitement automatique de cours magistraux <i>Salima Mdhaffar, Antoine Laurent et Yannick Estève</i>	419
Apprendre de la littérature scientifique : Les réseaux de signalisation en biologie systémique <i>Flavie Landomiel, Cathy Guérineau, Anubhav Gupta, Denis Maurel et Anne Poupon</i>	427
Détection des couples de termes translittérés à partir d'un corpus parallèle anglais-arabe <i>Wafa Neifar, Thierry Hamon, Pierre Zweigenbaum, Mariem Ellouze et Lamia Hadrich Belguith</i>	437
Utilisation d'une base de connaissances de spécialité et de sens commun pour la simplification de comptes-rendus radiologiques <i>Lionel Ramadier et Mathieu Lafourcade</i>	447
Algorithmes à base d'échantillonnage pour l'entraînement de modèles de langue neuronaux <i>Matthieu Labeau et Alexandre Allauzen</i>	455
Étude Expérimentale d'Extraction d'Information dans des Retranscriptions de Réunions <i>Pegah Alizadeh, Peggy Cellier, Thierry Charnois, Bruno Cremilleux et Albrecht Zimmermann</i>	465
Analyse morpho-syntaxique en présence d'alternance codique <i>José Carlos Rosales Núñez et Guillaume Wisniewski</i>	473

Simplification de schémas d'annotation : un aller sans retour ? <i>Cyril Grouin</i>	481
Apprentissage déséquilibré pour la détection des signaux de l'implication durable dans les conversations en parfumerie <i>Yizhe Wang, Damien Nouvel, Gaël Patin et Marguerite Leenhardt</i>	489
A comparative study of word embeddings and other features for lexical complexity detection in French <i>Aina Garí Soler, Marianna Apidianaki et Alexandre Allauzen</i>	499
Approche Hybride pour la translittération de l'Arabizi Algérien: Une enquête préliminaire <i>Imane Guellil, Azouaou Faical, Fodil Benali, Ala Eddine Hachani et Houda Saadane</i>	509
Lieu et nom de lieu, du texte vers la carte <i>Catherine Dominguès</i>	519
JeuxDeLiens: Word Embeddings and Path-Based Similarity for Entity Linking using the French JeuxDe-Mots Lexical Semantic Network. <i>Julien Plu, Kevin Cousot, Mathieu Lafourcade, Raphaël Troncy et Giuseppe Rizzo</i>	529
De l'usage réel des emojis à une prédiction de leurs catégories. <i>Gaël Guibon, Magalie Ochs et Patrice Bellot</i>	539
Transfert de ressources sémantiques pour l'analyse de sentiments au niveau des aspects. <i>Caroline Brun</i>	547
Apport des dépendances syntaxiques et des patrons séquentiels à l'extraction de relations. <i>Kata Gábor, Nadège Lechevrel, Isabelle Tellier, Davide Buscaldi, Haifa Zargayouna et Thierry Charnois</i>	557
Divergences entre annotations dans le projet UD et leur impact sur l'évaluation des performance d'étiquetage morpho-syntaxique. <i>Guillaume Wisniewski et François Yvon</i>	567
Annotation en Actes de Dialogue pour les Conversations en Ligne. <i>Robin Perrotin, Alexis Nasr et Jeremy Auguste</i>	577

Articles longs

