Traitement Automatique des Langues Naturelles (TALN) [1]

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles.

Volume 3 : Démonstrations

Pascal Denis, Natalia Grabar, Amel Fraisse, Rémi Cardon, Bernard Jacquemin, Eric Kergosien, Antonio Balvet (Éds.)

Lille, France, 28 juin au 2 juillet 2021

1. https://talnrecital2021.inria.fr/

Avec le soutien de

Soutiens institutionnels

Sponsors industriels

Partenaires « Argent »

Partenaires « Bronze »

Préface

Pour sa 28e édition, la conférence TALN s'est tenue pour la première fois de son histoire à Lille, et pour la seconde fois seulement dans la région des Hauts-de-France (après TALN 2009 à Senlis). Comme il en est devenu la tradition, TALN est une nouvelle fois organisée sous l'égide de l'ATALA conjointement avec sa conférence "soeur", RÉCITAL, dont c'est déjà la 23e édition.

Comme pour leurs éditions 2020, TALN 2021 et RÉCITAL 2021 ont à nouveau dû être "virtualisées" en raison de l'épidémie de Covid-19 qui a paralysé la France, l'Europe, et une bonne partie du monde. Ceci a considérablement compliqué son organisation et a conduit à la suppression de plusieurs événements originellement prévus dont le HackaTAL, la soirée gala, les événements sociaux, les promenades à Lille, la dégustation de la cuisine régionale, etc. Néanmoins, nous avons pu maintenir l'atelier Défi Fouilles de Textes (DEFT 2021), ainsi que non moins de 8 tutoriels différents. Nous remercions les organisateurs vaillants de DEFT et des tutoriels.

En lien avec cette actualité sanitaire, le thème choisi pour l'édition de TALN 2021 est "TAL et santé". Ce thème se reflète naturellement dans le programme de cette édition, puisqu'elle comprend une conférence invitée de Pierre Zweigenbaum sur le TAL médical, une session dédiée, et le traitement des cas cliniques comme tâche de DEFT 2021. Nous avons par ailleurs été très contents d'accueillir André Martins (professeur associé à l'Instituto Superior Técnico et VP recherche chez Unbabel, à Lisbonne au Portugal), comme second conférencier invité de cette édition.

Ces actes regroupent les articles des conférences TALN et RÉCITAL (volume 1 et 2, respectivement), ceux décrivant les démonstrations (volume 3), ceux issus de l'atelier DEFT 2021 (volume 4). Comme lors de la précédente édition de TALN 2020, un appel spécifique réservé aux résumés d'articles publiés dans des conférences internationales de premier plan fut également organisé. Ces résumés ont été versés dans le volume 1.

Pour TALN, un total de 58 articles a été soumis, soit exactement le même nombre que pour l'édition précédente. Parmi ceux-ci, 45 ont été sélectionnés, soit un taux d'acceptation de 77.6 %, dont 8 comme articles longs et 37 comme articles courts. Pour RÉCITAL, le nombre d'articles soumis fut de 16, en léger recul par rapport aux 22 soumissions de l'an dernier. 13 de ceux-ci ont été sélectionnés, soit un taux d'acceptation de 81.2 %.

Parmi les innovations de cette édition de TALN-RÉCITAL, nous avons rajouté une phase de discussion entre auteur(e)s et relecteures/relectrices, de manière à enrichir et fluidifier le processus de relecture et, on l'espère, à améliorer la séléction des articles et la plus-value des retours apportée aux auteur(e)s.

Nous sommes extrêmement reconnaissants à toutes les personnes qui ont participé aux différents comités scientifiques de ces conférences, à savoir :
— les responsables de domaine de TALN (voir page vi) ;
— les relectrices et relecteurs de TALN et RÉCITAL (voir page vi).
En outre, nous remercions chaleureusement l'ATALA, dont le comité permanent (le CPerm) assure la pérennité des TALN et RÉCITAL. Nous sommes également redevables à l'ensemble des membres du comité d'organisation (en particulier Antonio Balvet et Bernard Jacquemin), ainsi qu'aux personnes qui ont apporté leur soutien administratif et logistique

(en particulier Christine Yvoz) pour leur implication. Merci aussi à Yannick Parmentier qui nous a permis de produire ces actes et d'assurer la diffusion de ceux-ci sur HAL, l'ACL anthology et les archives TALN. Nous remercions aussi Onkar Pandit, Mariana Vargas et Nathalie Vauquier pour leur aide dans la maintenance du site web de la conférence et dans la configuration de la plate-forme `gather.town`.

Enfin, que soient aussi remerciés nos partenaires institutionnels et industriels pour leur soutien financier, en particulier : le CNRS, l'Inria, l'Université de Lille, les laboratoires CRIStAL, STL et GERIICO, l'ATALA et l'Afia, la GDLFLF, et les entreprises Schlumberger, ELRA, ERDIL, SINEQUA, ZENDOC.

Les présidentes et présidents de TALN : Pascal Denis et Natalia Grabar ;

Les présidentes et présidents de RÉCITAL : Amel Fraisse et Rémi Cardon.

Comités

Co-Président.e.s TALN

— Pascal Denis, MAGNET, Inria Lille & CRIStAL

— Natalia Grabar, STL, CNRS

Responsables de domaine

— Delphine Bernhard, LiLPA, Strasbourg

— Houda Bouamor, CMU Qatar

— Chloé Braud, IRIT, Toulouse

— Caroline Brun, NaverLabs, Grenoble

— Marie Candito, LLF, Paris

— Caio Corro, LISN, CNRS, Université Paris-Saclay

— Géraldine Damnati, Orange Labs, Lannion

— Maud Erhmann, EPFL, Suisse

— Cécile Fabre, CLLE, Toulouse

— Benoît Favre, TALEP, Marseille

— Thomas François, CENTAL, UCLouvain, Louvain-la-Neuve, Belgique

— Nuria Gala, LPL, Aix

— Philippe Langlais, DIRO, Montréal, Canada

— Philippe Muller, IRIT, Toulouse

— Alexis Nasr, TALEP, Marseille

— Magalie Ochs, LIS, Marseille

— Yannick Parmentier, LORIA, Nancy

— Tim van de Cruys, KUL, Leuven, Belgique

— Guillaume Wisniewski, LLF, Paris

Comité de lecture TALN

— Céline Alec, GREYC, Université de Caen-Normandie

— Alexandre Allauzen, LAMSADE, Université Paris-Dauphine

— Maxime Amblard, LORIA, Université de Lorraine

— Pascal Amsili, LATTICE, ILPGA, Université Sorbonne Nouvelle

— Loïc Barrault, University of Sheffield

— Patrice Bellot, Aix-Marseille Université – CNRS (LIS)

— Asma Ben Abacha, NLM/NIH, USA

— Laurent Besacier, Naver Labs Europe

— Yves Bestgen, F.R.S-FNRS et UCL

— Philippe Blache, LPL, CNRS

— Nathalie Camelin, LIUM, Le Mans Université

— Rémi Cardon, STL CNRS, Université de Lille

— Peggy Cellier, IRISA, INSA Rennes

— Thierry Charnois, LIPN, CNRS Université Sorbonne Paris Nord

— Vincent Claveau, IRISA, CNRS

— Maximin Coavoux, Université Grenoble Alpes, CNRS

— Mathieu Constant, ATILF, Université de Lorraine

— Benoit Crabbé, Université de Paris, LLF

— Béatrice Daille, LS2N, CNRS, Université de Nantes

— Mathieu Dehouck, CNRS, LATTICE

— Gaël Dias, Université de Normandie

— Patrick Drouin, OLST, Université de Montréal

— Emmanuelle Esperança-Rodier, LIG, Université Grenoble Alpes

— Dominique Estival, Western Sydney University

— Olivier Ferret, CEA List, Université Paris-Saclay

— Cyril Grouin, LISN, CNRS, Université Paris-Saclay

— Gaël Guibon, Télécom Paris et SNCF

— Olivier Hamon, Syllabs

— Thierry Hamon, Université Paris-Saclay, CNRS, LISN & Université Sorbonne Paris Nord

— Nabil Hathout, CLLE, CNRS

— Amir Hazem, LS2N, CNRS, Université de Nantes

— Nicolas Hernandez, LS2N, CNRS, Université de Nantes

— Stéphane Huet, LIA, Université d'Avignon

— Christine Jacquin, LS2N, CNRS, Université de Nantes

— Sylvain Kahane, Modyco, Université Paris Nanterre

— Mikaela Keller, MAGNET, Université Lille & CRIStAL

— Olivier Kraïf, LIDILEM, Université Grenoble Alpes

— Matthieu Labeau, Telecom Paris

— Éric Laporte, LIGM, Université Gustave Eiffel

— Gwénolé Lecorvé, Univ Rennes, CNRS, IRISA

— Benjamin Lecouteux, LIG, Université Grenoble Alpes

— Claire Lemaire, Lairdil, Université Paul Sabatier, Toulouse III ; LIG, Université Grenoble Alpes

— Yves Lepage, Université Waseda, Japon

— Cedric Lopez, EMVISTA

— Denis Maurel, Université de Tours, Lifat

— Anne-Lyse Minard, LLL, CNRS, Université d'Orléans

— Richard Moot, LIRMM, CNRS & Université de Montpellier

— Véronique Moriceau, IRIT, Université de Toulouse

— Emmanuel Morin, LS2N, CNRS, Université de Nantes

— Luka Nerima, LATL-CUI, Université de Genève

— Aurélie Névéol, LISN, CNRS, Université Paris-Saclay

— Jian-Yun Nie, Université de Montréal

— Damien Nouvel, ERTIM, INALCO

— Sylvain Pogodalla, LORIA, INRIA

— Jean-Philippe Prost, Aix-Marseille Université et Université de Montpellier

— Solen Quiniou, LS2N, CNRS, Université de Nantes

— Christian Raymond, IRISA, INSA Rennes

— Christian Retoré, LIRMM Univ Montpellier CNRS

— Sophie Rosset, LISN, CNRS, Université Paris-Saclay

— Didier Schwab, LIG, Université Grenoble Alpes

— Pascale Sébillot, IRISA, INSA Rennes

— Gilles Sérasset, LIG, Université Grenoble Alpes

— Ludovic Tanguy, CLLE, Université de Toulouse

— Xavier Tannier, Sorbonne Université, INSERM, LIMICS

— Andon Tchechmedjiev, EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales

— Charles Teissedre, SYNAPSE

— Juan-Manuel Torres-Moreno, Laboratoire Informatique d'Avignon / UA

— Nicolas Turenne, United International College, Chine

— François Yvon, LISN, CNRS, Université Paris-Saclay

— Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay

Co-Président.e.s RECITAL

— Amel Fraisse (GERIICO)

— Rémi Cardon (STL)

Comité de lecture RECITAL

— Jean-Yves Antoine (Université François Rabelais de Tours)

— Sonia Badene (Linagora IRIT)

— Rachel Bawden (INRIA)

— Johana Bodard (THIM/CHart EA4004 – Université Paris 8)

— Chloé Braud (IRIT – CNRS)

— Johanna Mayra Cordova (INALCO)

— Núria Gala (Aix-Marseille Université, LPL CNRS)

— Mahault Garnerin (Université Grenoble Alpes)

— Loïc Grobol (Lattice)

— William Havard (Université Grenoble Alpes)

— Laurine Huber (LORIA)

— Mikaela Keller (Université de Lille – INRIA)

— Yves Lepage (Waseda University)

— Anne-Laure Ligozat (LISN, CNRS, Université Paris-Saclay, ENSIIE)

— Damien Nouvel (INALCO)

— Patrick Paroubek (LISN, CNRS, Université Paris-Saclay)

— Thierry Poibeau (LaTTiCe-CNRS)

— Laurent Romary (INRIA & HUB-ISDL)

— Nicolas Turenne (INRA UPEM)

— Zheng Zhang (Schlumberger, AI Lab)

— Pierre Zweigenbaum (LISN, CNRS, Université Paris-Saclay)

Table des matières

ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs, multi-cibles, et Annotation d'Expressions Poly-lexicales **1**
Emmanuelle Esperança-Rodier, Francis Brunet-Manquat

Corpus EN-Istex : un corpus d'articles scientifiques annoté manuellement en entités nommées **6**
Enza Morale, Denis Maurel, Jeanne Villaneau, Jean-Yves Antoine

GECko+ : a Grammatical and Discourse Error Correction Tool **8**
Eduardo Calò, Léo Jacqmin, Thibo Rosemplatt, Maxime Amblard, Miguel Couceiro, Ajinkya Kulkarni

Outil Interactif et Évolutif pour l'Extraction d'Information dans des Documents Techniques **12**
Thiziri Belkacem, Charles Teissèdre

SIDRES : A Novel Annotation Tool For The Automatic Detection of Semantic Entities **15**
Julieta Murata, Rémy Carrette, Pierre Jourlin

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs, Multi-Cibles, et Annotation d'Expressions Polylexicales

Emmanuelle Esperança-Rodier[1] Francis Brunet-Manquat[1]
Univ. Grenoble Alpes, CNRS, Grenoble INP[1], LIG, 38000 Grenoble, France
prénom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Cette démonstration présente les avancées d'ACCOLÉ (Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs), qui en plus de proposer une gestion simplifiée des corpus et des typologies d'erreurs, l'annotation d'erreurs pour des corpus de traduction bilingues alignés, la collaboration et/ou supervision lors de l'annotation, la recherche de modèle d'erreurs dans les annotations, permet désormais d'annoter les Expressions Polylexicales (EPL) dans des textes monolingues en français, et d'accéder à l'annotation d'erreurs pour des corpus de traduction multi-cibles. Dans cet article, après un bref rappel des fonctionnalités d'ACCOLÉ, nous explicitons les fonctionnalités de chaque nouveauté.

ABSTRACT

ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora, Multi-Target corpora and Multi Word Expression Annotation.

This demonstration presents the recent advances in ACCOLÉ, which on top of offering simplified management of corpora and typologies of errors, annotation of errors in bilingual aligned corpora, collaboration and/or supervision during annotation, looking for error types in annotations, now permits to annotate Multi-Word Expressions (MWE) in French monolingual corpora, and to access error annotation for multi-target corpora. In this article, after reminding the regular features of ACCOLÉ, we will explain the features of each novelty.

MOTS-CLÉS : Annotations d'erreurs de Traduction Automatique, Annotation collaborative, Evaluation de la qualité de la TA, EPL

KEYWORDS: Annotations of translation errors, Collaborative annotation, Machine Translation Quality Assessment, MWE

[1] Institute of Engineering Univ. Grenoble Alpes

1 ACCOLÉ, une plateforme pour l'annotation d'erreurs

ACCOLÉ permet l'annotation manuelle des erreurs de traduction selon des critères linguistiques. L'idée sous-jacente est de pouvoir fournir à un utilisateur une aide dans le choix d'un système de TA à utiliser selon le contexte (compétences linguistiques et informatiques de l'utilisateur, connaissance du domaine du document source à traduire et la tâche pour laquelle il a besoin de traduire le document source.) Pour ce faire, ACCOLÉ doit permettre de détecter quels sont les phénomènes linguistiques qui ne sont pas traités correctement par le système de TA étudié. Les principales fonctionnalités de la plateforme ACCOLÉ sont la gestion simplifiée des corpus, des typologies d'erreurs, des annotateurs, etc. ; l'annotation d'erreurs ; la collaboration et/ou supervision lors de l'annotation ; la recherche de modèles d'erreurs (type d'erreurs dans un premier temps, patrons morphosyntaxiques ultérieurement) dans les annotations. Nous avons privilégié un accès simple à l'outil ainsi qu'au corpus. La plateforme ACCOLÉ est donc disponible en ligne (http://lig-accole.imag.fr,) depuis un navigateur et ne nécessite aucune installation spécifique.
Un projet d'annotation renvoie à une tâche d'annotation, en créant un couple associant un corpus et une typologie d'annotation. Ainsi, un même corpus pourra être associé à plusieurs typologies sous forme de plusieurs projets d'annotation. Le corpus ne sera chargé qu'une fois sur la plateforme. Les annotateurs ainsi que les superviseurs sont associés aux projets qu'ils doivent annoter par le responsable du projet. Les typologies d'erreur sont également gérées par le responsable du projet. Un type d'erreur est composé d'un nom, d'une catégorie (facultative), d'une sous-catégorie (facultative) et d'un code (raccourci clavier pouvant être utilisé lors de l'annotation).
L'annotation se fait en deux étapes. La première étape consiste à sélectionner, à l'aide de la souris, des mots dans la phrase source, et de leur équivalent dans la phrase cible, présentant une erreur de traduction. Il est possible de sélectionner des mots disjoints dans la source et dans la cible. Dans le cas de mots non traduits (omission), il faut sélectionner l'espace dans la cible, à l'endroit où le ou les mots sources aurait dû être traduits. Dans le cas d'addition, il faut sélectionner l'espace entre les mots sources, correspondant à la position du ou des mots qui ont été ajoutés dans la cible entre les traductions de ces mots sources. La seconde étape consiste à choisir le type d'erreurs soit à l'aide de la souris, soit à l'aide des raccourcis clavier, à associer au couple des mots sources/cibles préalablement sélectionnés.
ACCOLÉ est une plateforme d'annotation collaborative, elle est donc dotée de deux mécanismes d'aide à l'annotation. Le premier est un mécanisme de supervision permettant à un responsable de contrôler l'avancée de la tâche. Ce mécanisme encourage surtout la communication entre superviseur et annotateur par la possibilité de créer des fils de discussion pour un couple de phrase source/cible précis. La supervision autorise la demande de précisions sur un type d'erreurs, de pointer une erreur d'annotation, etc.. Le second mécanisme est le mécanisme collaboratif qui permet aux annotateurs de communiquer autour d'un couple phrase source/cible précis. Ce mécanisme est une option à activer dans le projet. Du fait de son aspect collaboratif, ACCOLÉ répond aux problèmes d'accord inter-annotateurs (Popović, 2018).

2 ACCOLÉ, nouvelles fonctionnalités

2.1 Annotation d'erreurs multi-cibles

Pour l'analyse de la qualité de systèmes de Traduction Automatique Neuronale en traduction simultanée ou après complétion de la phrase entière - online & offline NMT - (Elbayad et al., 2020), ACCOLÉ s'est étoffée de l'annotation de plusieurs hypothèses de traduction correspondant à une seule phrase source et de l'intégrer d'une phrase de référence.
L'annotation d'erreurs sur un corpus multi-cibles se déroule de la même façon que pour un corpus mono-cible. L'annotateur sélectionne à l'aide de la souris le couple d'occurrence source/cible 1 source/cible 2, source /cible 3… présentant une erreur de traduction. La seconde étape consiste à choisir le type d'erreurs soit à l'aide de la souris, soit à l'aide des raccourcis clavier, à associer au couple des mots sources/cibles préalablement sélectionnés. En plus de la source, l'annotateur a accès à une traduction de référence, comme le montre la Figure 1 ci-dessous.

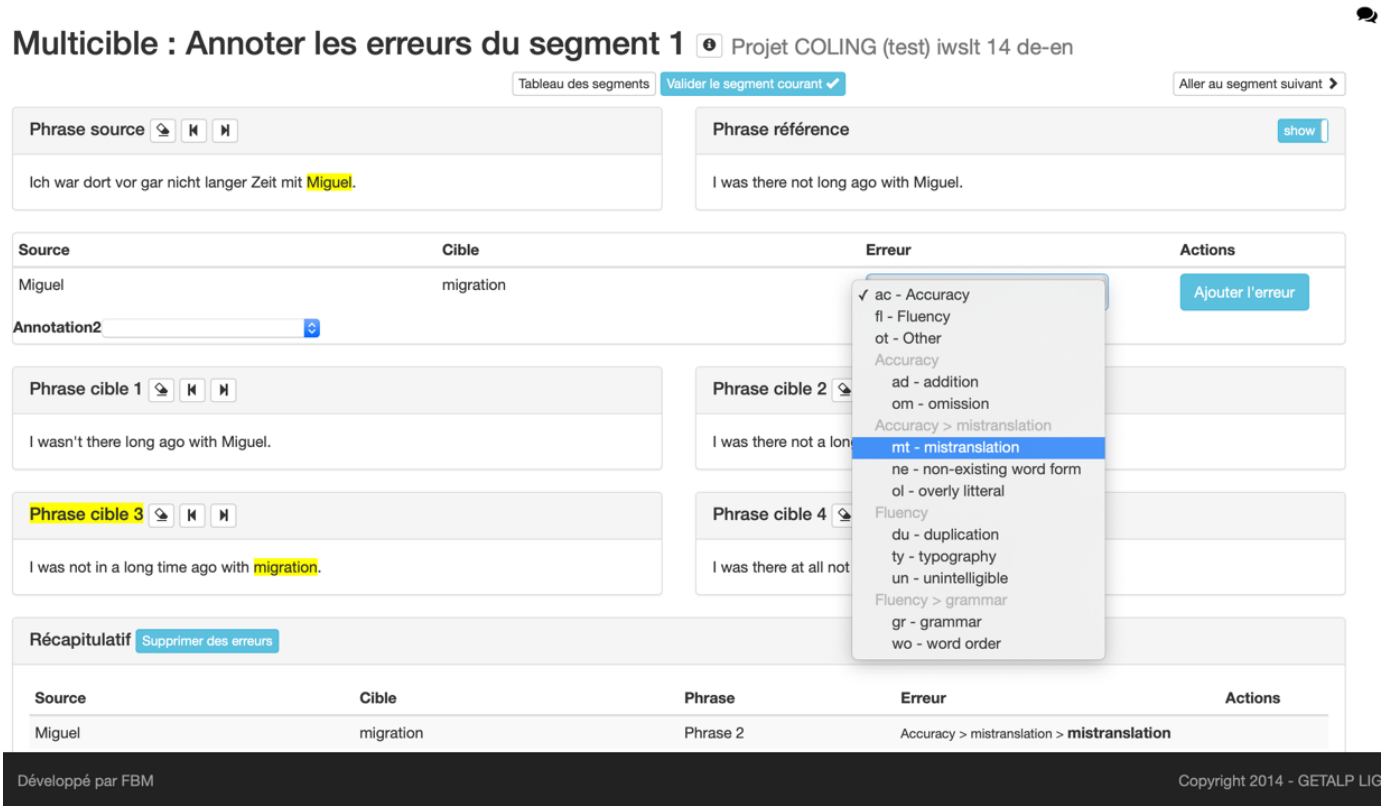

FIGURE 1 : Annotation d'une erreur sur la plateforme ACCOLÉ avec la typologie DQF-MQM (Lommel et al., 2018) dans un corpus multi-cible avec référence.

2.2 Annotation d'Expressions Polylexicales

Nous avons adapté ACCOLÉ pour l'annotation d'Expressions Polylexicales (EPL) comme l'illustre la Figure 2. La typologie de types d'EPL intégrée à notre plateforme est telle que définie dans le

travail de Tutin & al. (2017), composée de 9 types : Collocations, Mots Fonctionnels, Formules de Routine, Entités nommées, Phrasèmes complets, Pragmatèmes, Proverbes, Collocations fortes et enfin Termes Complexes. Chaque EPL est également annotée en partie du discours.
Nous avons envisagé que le corpus annoté soit monolingue ou bien bilingue. Toutefois, nous préférons la possibilité d'annoter la source en EPL de manière monolingue, de même que la cible. Si le corpus possède une traduction alignée du texte, alors il est possible d'annoter, à la fois dans la source et dans la cible, l'erreur repérée entre une EPL et sa traduction, afin de faire correspondre et comparer les annotations faites en première étape monolingue
Afin de faciliter la tâche d'annotation, un dictionnaire monolingue français d'EPL a été ajouté à ACCOLÉ, ainsi qu'un pré-traitement basé sur l'analyse syntaxique (Coavoux et Crabbé, 2017). Ainsi, ACCOLÉ permet d'annoter des EPL soit manuellement, en sélectionnant des mots à l'aide de la souris et en leur assignant un type, soit sur proposition de l'interface utilisant de manière automatique le dictionnaire et le pré-traitement, proposition qui sera à valider par l'annotateur.

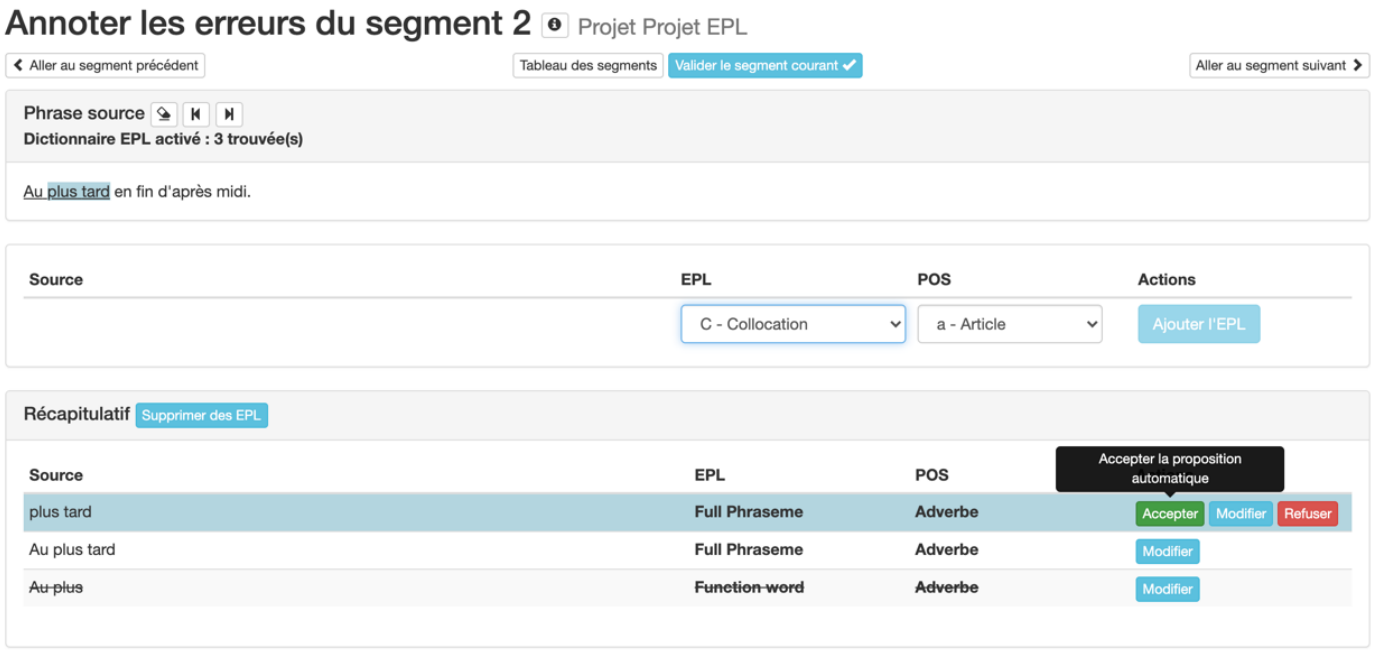

FIGURE 2 : Annotation d'un segment en EPL avec proposition automatique d'annotations

3 Données disponibles

ACCOLÉ propose 3 typologies d'erreurs, celle de Vilar et al. (2006), deux autres issues de MQM-DQF (Lommel, 2018) et 1 typologie d'annotation des EPL (Tutin et al, 2017) ainsi que 14 corpus FR-GB, 4 corpus monolingues FR/GB et 7 corpus GB-DE (allant des nouvelles journalistiques, à des documents techniques, des brevets, des extraits du BTEC (Basic Travel Expression Corpus) jusqu'à des documents sur le climat ou des textes médicaux) pour un total de 25 corpus (+66,6% en un an), ayant permis la création de 30 projets (+58%). Ceux-ci correspondent à 9 585 phrases (+40,5 %), 184 786 mots sources (+37,6%), 266 078 mots cibles (+132%), pour 34 558 annotations réalisées par 12 annotateurs natifs soit anglais, allemand ou français (+47%). Ces corpus sont

structurés selon les SNODEs (Boitet et al., 1988) et sont disponibles sur demande au format XML ou JSON. Une fonction permet de rechercher dans ces corpus les types d'erreurs. Au moment de la rédaction, nous continuons de travailler sur la recherche de modèle d'erreurs et plusieurs projets d'annotation sont en cours.

Références

BOITET C. ET ZAHARIN Y. (1988). Representation trees and string- tree correspondences. In *Proceedings of international Conference on Computational Linguistics* COLING-88, 59-64.

COAVOUX M. ET CRABBÉ B. (2017). Représentation et analyse automatique des discontinuités syntaxiques dans les corpus arborés en constituants du français. *Actes de la 24e conférence sur le Traitement Automatique des Langues Naturelles, Jun 2017, Orléans, France. pp.77-92*

ELBAYAD M., USTASZEWSKI M., ESPERANÇA-RODIER E., BRUNET-MANQUAT F., VERBEEK, J. ET BESACIER L. (2020). Online Versus Offline NMT Quality: An In-depth Analysis on English–German and German–English. *Accepté à COLING 2020.*

LOMMEL A., ET ALAN K. M. (2018). Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). *13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers). Vol. 2.*

POPOVIĆ, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. *Moorkens J., Castilho S., Gaspari F., Doherty S. (eds) Translation Quality Assessment. Machine Translation: Technologies and Applications, vol 1. Springer.*

TUTIN A., ESPERANÇA-RODIER E. (2017). La difficile identification des expressions polylexicales dans les textes : critères de décision et annotation. "*La phraséologie française : débats théoriques et dimensions appliquées (didactique, traduction et traitement informatique)", Sep 2017, Arras, France.*

VILAR D., XU J., D'HARO L.F. ET AL. (2006). Error analysis of statistical machine translation output. *5th International Conference on Language Resources and Evaluation.* 97-702.

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Corpus EN-ISTEX : un corpus d'articles scientifiques annoté manuellement en entités nommées

Enza Morale[1], Denis Maurel[2],
Jeanne Villaneau[3], Jean-Yves Antoine[2]

(1) Inist, CNRS, Nancy, France

(2) Université de Tours, Lifat, Tours, France

(3) Ensibs, Irisa, Lorient, France

enza.morale@inist.fr, denis.maurel@univ-tours.fr,
Jean-Yves.Antoine@univ-tours.fr, jeanne.villaneau@univ-ubs.fr

RESUME

Nous présentons ici une nouvelle ressource libre : le corpus EN-ISTEX, un corpus de deux cents articles scientifiques annotés manuellement en entités nommées. Ces articles ont été extraits des deux éditeurs scientifiques les plus importants de la plateforme ISTEX. Tous les domaines sont concernés, même si les sciences dites *dures*, en particulier les sciences du vivant et de la santé, sont prépondérantes.

Parmi ceux-ci vingt articles ont été multi-annotés afin de vérifier l'adéquation du guide d'annotation et la fiabilité de l'annotation. L'accord inter annotateurs sur ces vingt textes s'élève à 91 %.

ABSTRACT

ISTEX-EN Corpus: a scientific paper corpus manually annotated in named entities
We present here a new free resource: the EN-ISTEX Corpus, a corpus of two hundred scientific papers manually annotated in named entities. These papers have been extracted from the two more representative scientific publishers of ISTEX platform. All fields are concerned, even if the so-called hard sciences, in particular the life sciences and health, are predominant.

Among these, twenty papers were multi-annotated in order to verify the adequacy of the annotation guide and the reliability of the annotation. The inter-annotator agreement on these twenty texts amounts to 91%.

MOTS-CLES : corpus annoté, entités nommées, ressource libre, articles scientifiques, accord inter annotateurs.

KEYWORDS: annotated corpus, named entities, free resource, scientific papers, inter annotator agreement.

---

1 Introduction

Le projet investissement d'avenir ISTEX[1] avait pour but la constitution d'une bibliothèque d'articles scientifiques disponibles en libre accès pour les acteurs de la recherche en France. Cette bibliothèque numérique comporte aujourd'hui « *23 millions de documents provenant de 30 corpus de littérature scientifique dans toutes les disciplines* »[2]. Pour améliorer la consultation de la plateforme ISTEX, des services à valeurs ajoutées ont été développés et sont disponibles via l'API d'ISTEX. Parmi ceux-ci, un service permet une interrogation de la base via les entités nommées (noms propres, dates et références). Ce service a bien sûr été testé en interne, mais l'idée est venue de constituer manuellement un corpus annoté en entités nommées avec un accord inter annotateurs, utilisable comme corpus d'apprentissage pour la création ou l'amélioration d'outils de détection d'entités nommées. Il s'agit d'une nouvelle ressource libre (sous licence ouverte Etalab[3]), disponible à l'URL https://corpus-gold.corpus.istex.fr/.

Ce corpus contient deux cents articles, issus des éditeurs Wiley et Elsevier, sélectionnés à partir des catégories scientifiques Science-Metrix de niveau 1, proportionnellement à l'ensemble constitué de ces deux éditeurs dans le fonds Istex. Parmi ceux-ci, vingt articles ont fait l'objet d'une annotation multiple avec calcul de l'accord inter annotateurs et choix collégial de la bonne annotation. Cet accord, calculé par le *alpha* de Krippendorff, est très bon, puisqu'il atteint 91 %. Le travail sur ces vingt articles a permis la formation des annotateurs qui ont ainsi acquis une capacité à annoter les textes de façon similaire. De ce fait, l'ensemble du corpus EN-ISTEX peut être considéré comme un *gold standard*. L'interprétation du *alpha* porte toujours à débat, mais pour un coefficient alpha de cet ordre, cette ressource peut être qualifiée d'une très bonne fiabilité.

2 Démonstration

Nous présenterons :

- la constitution du corpus à partir des catégories scientifiques Scopus ;
- la campagne d'annotation ;
- le guide d'annotation (téléchargeable sur le site) avec différents exemples, en particulier quelques points qui ont donné lieu à discussion ;
- L'accès au corpus annoté via le site de publication des corpus Istex : http://data.istex.fr/.

---

1 Le projet ISTEX (ANR-10-IDEX-0004-02) s'est déroulé d'avril 2012 à décembre 2018.

2 D'après le site https://www.istex.fr/, consulté le 26/11/2020.

3 https://www.etalab.gouv.fr/licence-ouverte-open-licence

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

GECko+: a Grammatical and Discourse Error Correction Tool

Eduardo Calò[1*] Léo Jacqmin[1*] Thibo Rosemplatt[1*]
Maxime Amblard[12] Miguel Couceiro[12] Ajinkya Kulkarni[2]
(1) IDMC, Université de Lorraine, F-54000 Nancy, France
(2) Université de Lorraine, CNRS, Inria N.G.E., LORIA, F-54000, France
{eduardo.calo6,leo.jacqmin8,thibo.rosemplatt3}@etu.univ-lorraine.fr,
{maxime.amblard,miguel.couceiro,ajinkya.kulkarni}@loria.fr

RÉSUMÉ

Nous présentons GECko+, un assistant d'écriture pour l'anglais qui corrige des erreurs au niveau de la phrase et du discours. Il se base sur deux modèles état de l'art pour la correction grammaticale et pour la réorganisation de phrases. GECko+ est disponible en ligne sous la forme d'une application web qui implémente une chaîne de traitement assemblant ces deux modèles.

ABSTRACT

GECko+ : a Grammatical and Discourse Error Correction Tool

We introduce GECko+, a web-based writing assistance tool for English that corrects errors both at the sentence and at the discourse level. It is based on two state-of-the-art models for grammar error correction and sentence ordering. GECko+ is available online as a web application that implements a pipeline combining the two models.

MOTS-CLÉS : assistant d'écriture, correction grammaticale, analyse de discours.

KEYWORDS: writing assistant tool, grammatical error correction, discourse analysis.

1 Introduction

While most people can write, few would boast they never produce spelling and grammar mistakes, let alone systematically write coherent prose and express ideas clearly. Natural language processing (NLP) techniques have the potential to help in that regard. In particular, such technologies can have a beneficial impact on two issues related to the way we write.

First, NLP techniques can help us alleviate language-related discrimination (Papakyriakopoulos *et al.*, 2020), that occurs, e.g., in the professional world where job applications are rejected simply due to the quality of one's writing. Additionally, errorful writing is poorly perceived in social contexts and is often synonymous with barriers.

Second, those who are already proficient in writing can benefit from these techniques to improve the quality of their prose. This aspect applies to journalists, business-persons, college students, and teachers alike. These individuals are often required to write lengthy reports. The frequency with which these reports are produced is such that topological or consistency errors can occur. As a result,

∗. These authors contributed equally to this work.

their message may not be delivered as intended.

To address these issues, we propose a digital writing assistance tool for English that we call GECko+ that uses existing state-of-the-art models to tackle both sentence-level mistakes and discourse incoherence. To correct spelling and grammar mistakes, we use GECToR (Omelianchuk *et al.*, 2020), a grammatical error correction (GEC) model developed by the well-known Grammarly [1]. For tackling discourse incoherence, we make use of a sentence ordering model [2] (Prabhumoye *et al.*, 2020) based on Google's BERT (Devlin *et al.*, 2019). We created a web interface that users can access to correct paragraphs of text in English [3]. The code is publicly available on GitHub [4].

2 Background

GEC in NLP encompasses any sort of modifications made to automatically correct an errorful sentence. This includes spelling, punctuation, grammar, and word choice errors. Given a potentially errorful sentence or short piece of text as input, a GEC system is expected to output a corrected version of that text. We have reviewed several approaches to correct sentences individually (Chollampatt & Ng, 2018; Junczys-Dowmunt *et al.*, 2018).

However, language does not simply consist of individual, independent sentences that are added one after the other, but rather forms a coherent whole composed of interconnected sentences. This coherent whole is commonly referred to as discourse. The area of NLP concerned with how sentences fit together is called discourse coherence or discourse analysis (Jurafsky & Martin, 2009). Discourse analysis encompasses many different aspects and can be very fine-grained. One of these aspects is sentence ordering, whose goal is to arrange sentences of a given text in the correct order, i.e., in a coherent manner.

3 Description of the Tool

GECko+ combines two state-of-the-art models into a single pipeline. To tackle sentence-wise errors, it employs GECToR (Omelianchuk *et al.*, 2020), which treats GEC as a sequence tagging task, relying on a Transformer-based encoder. To address discourse coherence, it utilizes a sentence ordering model (Prabhumoye *et al.*, 2020), which predicts the relative ordering between pairs of sentences from an input list of sentences. The reordering task is treated as a constraint learning task. The pipeline is shown in Figure 1.

As the diagram shows, the text given as input by the user gets segmented into sentences. After the segmentation, we obtain a list S of sentences, whose length ranges from one to n. Then, GECToR is applied to each sentence s_i in S, in order to perform sentence-wise error correction. Each sentence is iteratively processed by the model to ensure that all interdependent errors get corrected. As a result, the n sentences that constitute S are now free of grammatical errors. Subsequently, if $n = 1$, the single corrected sentence is directly output to the user. Conversely, if S contains more than one

1. https://github.com/grammarly/gector
2. https://github.com/shrimai/Topological-Sort-for-Sentence-Ordering
3. https://gecko-app.azurewebsites.net/
4. https://github.com/psawa/gecko-app

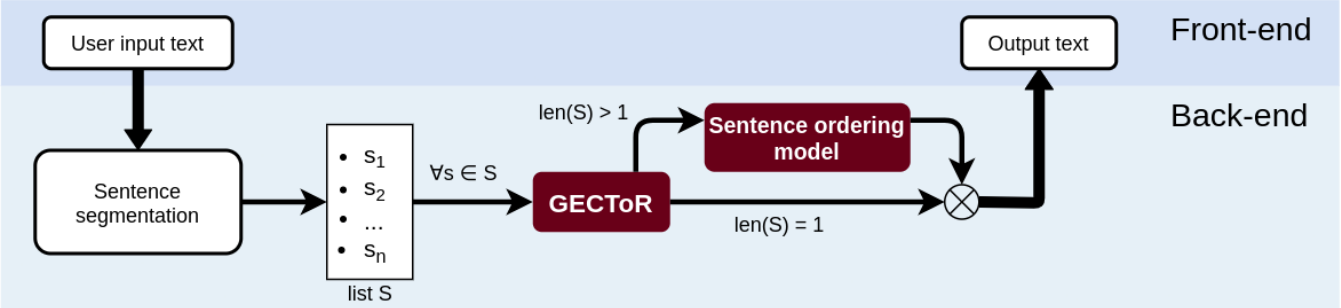

FIGURE 1 – GECko+'s pipeline.

element, the potentially unordered list of sentences will be given as input to the sentence ordering model. Once the sentences are ordered, the output is displayed to the user.

GECko+ employs a simple but effective color code to highlight mistakes. Changes are highlighted token-wise : deletions are underlined in red, modifications in blue, and additions in green. Currently there is no explicit indication of how sentences have been reordered. Ideally, a user should be able to visualize which sentences were swapped. We leave it for future work. Refer to Figure 2 for GECko+'s interface with an example sentence containing various spelling, grammar, and discourse mistakes.

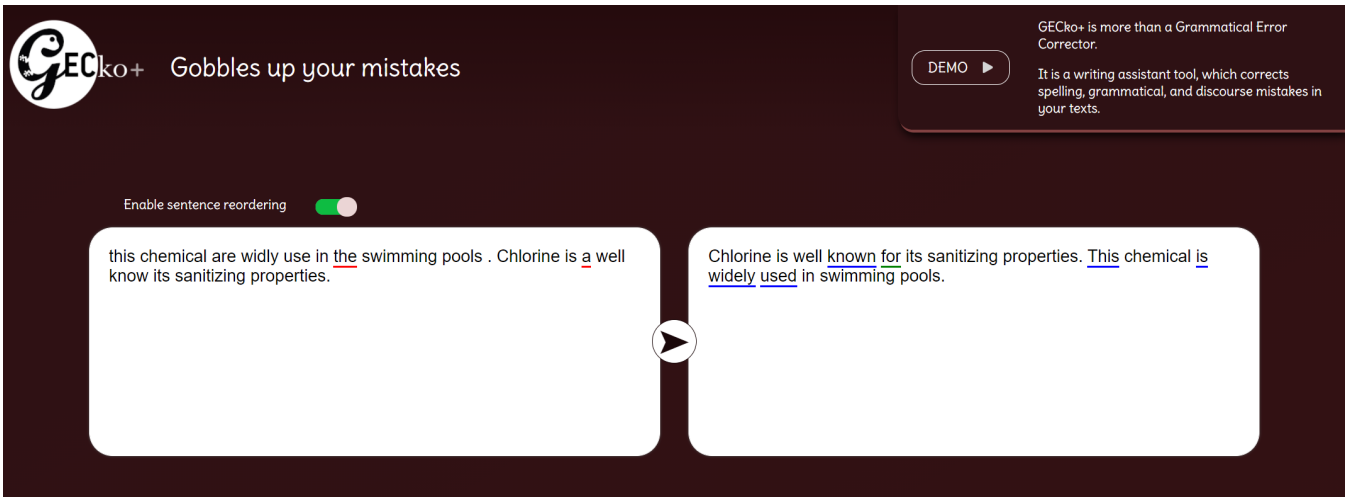

FIGURE 2 – GECko+'s interface.

4 Evaluation

The results for GECToR have been reported on CoNLL-2014 test set (Ng *et al.*, 2014) with the M^2 Scorer (Dahlmeier & Ng, 2012) and on BEA-2019 development and test sets (Bryant *et al.*, 2019) with ERRANT (Bryant *et al.*, 2017). For single models, they achieved state-of-the-art performance with an XLNeT-based model, which we use for our application, obtaining $F_{0.5} = 65.3$ on CoNLL-2014 (test) and $F_{0.5} = 72.4$ on BEA-2019 (test). The sentence ordering model was evaluated across several datasets using multiple metrics along with a human evaluation. The BERT-based approach scored higher than the previous state-of-the-art method on all metrics, obtaining a Sentence Accuracy of 61.48 on the NIPS dataset. Refer to (Prabhumoye *et al.*, 2020) for a detailed description of the results.

Références

BRYANT C., FELICE M., ANDERSEN Ø. E. & BRISCOE T. (2019). The bea-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 52–75 : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4406.

BRYANT C., FELICE M. & BRISCOE T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 793–805, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1074.

CHOLLAMPATT S. & NG H. T. (2018). A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, p. 5755–5762, New Orleans, Louisiana USA.

DAHLMEIER D. & NG H. T. (2012). Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 568–572, Montréal, Canada : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

JUNCZYS-DOWMUNT M., GRUNDKIEWICZ R., GUHA S. & HEAFIELD K. (2018). Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 595–606, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1055.

JURAFSKY D. & MARTIN J. H. (2009). *Speech and Language Processing (2nd Edition)*. USA : Prentice-Hall, Inc.

NG H. T., WU S. M., BRISCOE T., HADIWINOTO C., SUSANTO R. H. & BRYANT C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning : Shared Task*, p. 1–14, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/W14-1701.

OMELIANCHUK K., ATRASEVYCH V., CHERNODUB A. & SKURZHANSKYI O. (2020). GECToR – grammatical error correction : Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 163–170, Seattle, WA, USA. Online : Association for Computational Linguistics.

PAPAKYRIAKOPOULOS O., HEGELICH S., SERRANO J. C. M. & MARCO F. (2020). Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 446–457, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3351095.3372843.

PRABHUMOYE S., SALAKHUTDINOV R. & BLACK A. W. (2020). Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2783–2792, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.248.

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Outil Interactif et Évolutif pour l'Extraction d'Information dans des Documents Techniques

Thiziri Belkacem Charles Teissèdre
Synapse Développement, 7 Boulevard de la gare, 31500 Toulouse, France
thiziri.belkacem@synapse-fr.com; charles.teissedre@synapse-fr.com

RÉSUMÉ

L'accès à l'information dans la documentation technique est une application particulière et complexe du traitement du langage naturel et de la recherche d'information. La difficulté tient aux contraintes propres des langages métier spécialisés et semi-contrôlés. Dans ce document, nous proposons un outil d'accès à l'information dans différents types de documents. Notre solution exploite conjointement la structure organisationnelle des documents et leur contenu informationnel, pour extraire des informations métier dans des différents corpus. Nous proposons un système basé sur des interactions expert-machine dans un cycle d'amélioration continu des modèles d'extraction. Notre approche exploite des modèles d'apprentissage à faible supervision ne nécessitant pas d'expertise en ingénierie des langues. Notre système intègre l'utilisateur dans le processus de qualification de l'information et permet de guider son apprentissage, afin de rendre ses modèles plus performants au fil du temps.

ABSTRACT

Interactive and Evolutive Tool for Information Extraction in Technical Documents

Information access in technical documentation is a particular and complex application of natural language processing and information retrieval. The difficulty lies in the specific constraints of specialized and semi-controlled specialized business languages. In this paper, we propose a tool for accessing information in different types of documents. Our solution jointly exploits the organizational structure of documents and their information content to extract specific pieces of business information from different corpora. We propose a system based on expert-machine interactions in a cycle of continuous improvement of extraction models. Our approach exploits weakly supervised learning models that do not require expertise in language engineering. Our system integrates the user in the information qualification process and allows guiding the user's learning, in order to make the models more efficient over time.

MOTS-CLÉS : Extraction d'Information, Document Technique, Modèle Évolutif..

KEYWORDS: Information Extraction, Technical Document, Evolutive Model..

1 Description

L'accès à l'information dans des corpus de textes de forte technicité est rendu ardue du fait des contraintes propres aux langages métier spécialisés et régulièrement utilisés dans la documentation des industries. Dans cette dernière, les informations et connaissances pertinentes peuvent se présenter

sous différents formats et avoir une certaine régularité ou norme, dépendante du contenu et propre au domaine.

Indépendamment du domaine, différentes applications de traitement du langage naturel et de recherche d'information nécessitent une indexation des séquences informationnelles. Dans des domaines de spécialité, agréger des données d'entraînement permettant de repérer et indexer de telles séquences implique de mobiliser des experts métier en mesure d'analyser et qualifier le contenu des documents. Or la disponibilité des experts - une ressource rare - constitue un frein à la constitution de telles ressources.

Du fait que les modèles de langue génériques sont généralement moins performants dans les domaines spécifiques (Salloum *et al.*, 2020; Torfi *et al.*, 2020), adapter ces modèles à des domaines de spécialité se heurte ainsi à la difficulté de trouver des données qualifiées en volume suffisant (Kadhim, 2019; Chawla *et al.*, 2004). Comme les modèles de langue pré-entraînés sur des corpus de langue tous domaines peinent à analyser un vocabulaire et une langue régis par des contraintes spécifiques, propres à un métier ou à une organisation, et différentes de celles des langues naturelles (Ramponi & Plank, 2020; Ji *et al.*, 2021; Pathak *et al.*, 2020), il est donc impératif de trouver un moyen de préparer des données qualifiées en sollicitant des experts de la façon la plus parcimonieuse et optimale possible. L'outil que nous avons développé repose sur une approche hybride. Cette dernière consistant à analyser conjointement le contenu informationnel véhiculé par le texte, d'un côté, et par la structure organisationnelle du texte lui-même, d'un autre côté, permet de construire progressivement des modèles performants d'accès à l'information en impliquant un effort minimal de l'expert. La structure organisationnelle de la documentation est souvent riche en informations, en particulier dans les domaines industriels où la documentation est fortement normalisée. Cette structure traduit une hiérarchisation et une organisation des informations contenues dans les documents selon une logique métier, et qui doit être représentée dans un format exploitable pour des tâches d'extraction et de recherche d'information. Nous proposons une approche originale s'appuyant sur un cycle d'amélioration continue (lifelong learning) (Field, 2000). Afin d'affiner progressivement les prédictions proposées, ces modèles ont été intégrés dans un outil interactif permettant de mettre en oeuvre une forme de supervision faible par des experts dans un cycle itératif de validation des extractions, où la machine apprend à faire des prédictions de plus en plus précises et couvrantes à mesure que de nouveaux exemples et contre-exemples sont collectés. Les exemples (ou contre exemples) évalués peuvent ainsi servir de nouvelles données d'entraînement pour renforcer la faculté de prédiction du système.

L'approche que nous avons mis en oeuvre exploite différents axes de lecture : l'axe de la structure de la documentation (la façon dont les informations sont organisées et découpées à l'intérieur des documents) et celui de la logique métier (information dépendante du domaine d'utilisation). Nous avons utilisé des modèles d'apprentissage supervisés permettant de construire des représentations liées à la fois à la mise en forme des documents et à leur contenu informationnel. Ainsi, l'interface du système permet à la fois d'accéder aux informations pertinentes dans la documentation, corriger les prédictions du système lorsque celui-ci se trompe, entraîner de nouveaux modèles ou ré-entraîner des modèles existants et enfin, construire des données qualifiée pour l'entraînement de nouveaux modèles.

Références

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CHAWLA N. V., JAPKOWICZ N. & KOTCZ A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, **6**(1), 1–6.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FIELD J. (2000). *Lifelong learning and the new educational order*. ERIC.

JI S., HÖLTTÄ M. & MARTTINEN P. (2021). Does the magic of bert apply to medical code assignment ? a quantitative study. *arXiv preprint arXiv :2103.06511*.

KADHIM A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, **52**(1), 273–292.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

PATHAK A. R., AGARWAL B., PANDEY M. & RAUTARAY S. (2020). Application of deep learning approaches for sentiment analysis. In *Deep Learning-Based Approaches for Sentiment Analysis*, p. 1–31. Springer.

RAMPONI A. & PLANK B. (2020). Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6838–6855.

SALLOUM S. A., KHAN R. & SHAALAN K. (2020). A survey of semantic analysis approaches. In *Joint European-US Workshop on Applications of Invariance in Computer Vision*, p. 61–70 : Springer.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

TORFI A., SHIRVANI R. A., KENESHLOO Y., TAVVAF N. & FOX E. A. (2020). Natural language processing advancements by deep learning : A survey. *arXiv preprint arXiv :2003.01200*.

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

SIDRES : A Novel Annotation Tool For The Automatic Detection Of Semantic Entities

Julieta Murata[1] Rémy Carrette[1, 2] Pierre Jourlin[2]
(1) SATT Sud-Est, Le Silo, 35 Quai du Lazaret, CS 70545 13304 Marseille Cedex 02, France
(2) Laboratoire d'Informatique - Avignon Université, 339 Chemin des Meinajaries, 84000 Avignon, France
julietamurata@gmail.com, remy.carrette@alumni.univ-avignon.fr, pierre.jourlin@univ-avignon.fr

RÉSUMÉ

Nous présentons un nouvel outil d'annotation nommé SIDRES (Système Interactif de Détection et de Reconnaissance d'Entités Sémantiques). SIDRES fournit un environnement d'annotation pour la classification d'unités textuelles à partir de catégories *ad hoc*. Ces catégories peuvent être associées à des contextes comme un moyen de désambiguïsation d'unités identiques appartenant à des catégories différentes. SIDRES a été développé dans le cadre d'un partenariat industriel visant à effectuer un transfert de technologie du laboratoire de recherche publique vers des acteurs de l'industrie.

ABSTRACT

We present a novel annotation tool called SIDRES (Système Interactif de Détection et de Reconnaissance d'Entités Sémantiques [Interactive System for the Detection and Identification of Semantic Entities]). SIDRES provides an annotation environment for classifying text units through *ad hoc* categories. These categories can be coupled with contexts, so as to provide a means for the disambiguation of formally identical units assigned to different categories. SIDRES was developed as part of an industrial partnership between the LIA (Laboratoire d'Informatique d'Avignon [Research Institute of Informatics at the University of Avignon]) and a French company in the eHealth sector. This partnership was created within the framework of a technology-transfer project promoted by the SATT Sud-Est, whose core mission is bringing together industry and research institutions.

MOTS-CLÉS : rapports cliniques, annotation, grammaire ambigüe locale

KEYWORDS: medical reports, annotation, locally ambiguous grammar

1 Theoretical Aspects

SIDRES relies on a particular text data structure called "confusion tree". Confusion trees allow to define locally ambiguous grammars capable of representing multi-word expressions, determine their boundaries, and provide means for disambiguation as well as group these expressions according to different syntactic or semantic categories based on their linguistic context. This type

of structure is exploited by an algorithm in the line of Tomita's Generalized LR (Tomita, 1984). The algorithm allows to extract and disambiguate concurrent terminology in natural language texts.

The example below shows the term "Paris" being categorized as both a location (the capital city of France) and a person (Paris Hilton). Subsequent linguistic contexts allow for the disambiguation of these terms.

Paris (Location : 111 ; Person : 15)
|____ de Paris (Location : 47, Person : 5)
| |____ville de Paris (Location : 47)
|____ Paris lance (Location : 20, Person : 3)
| |____ Paris lance un (Location : 10, Person : 1)
| |____ Paris lance un audit (Location : 5)
|____ Paris, (Location : 100 ; Person : 12)
|____ Paris, considérée (Location : 80 ; Person : 5)
|____ Paris, considérée comme (Location : 79 ; Person : 5)
|____ Paris, considérée comme une (Location : 75 ; Person : 4)
|____ Paris, considérée comme une jet-setteuse (Person : 2)

2 Functionalities

2.1 Interface Design, Categories, and Contexts

SIDRES is coded in Python3 and uses the GTK4 framework. The interface design has a two-panel division: the left-hand panel provides for the creation of annotation categories, and the right-hand panel consists of a text visualization window. In order for the annotation corpus to be displayed on the right, the corpus must first be loaded either by connecting to a local database or by selecting a text file from the local file system. Annotators can then manually define and edit a set of categories of their choice (e.g.: <person>, <location>) according to their own annotation model, and assign a distinct color to each category. Upon manually selecting and right-clicking on textual units, the user can simply add the unit to the intended category. If a unit is annotated under more than one category (e.g.: the term “Paris” is a <person> and a <location>), users can select a linguistic context so as to provide a means of disambiguation. At any point in the annotation process, users can save and export their work in a JSON file, and load it up again.

2.2 Other Features

To satisfy the goals of annotation management, the upper menu allows for further functions, namely: (1) a display function that presents the list of categories and a cumulative list of annotated units; (2) a research function to find identical units that have been assigned to different categories; (3) basic statistics on the annotated units (total number of units, categories, distribution graphs).

2.3 Use Examples in the Medical Domain: the Negation of Signs and Symptoms, and the Negation of Medical History in Medical Reports in French

We present two case examples of corpus annotation of medical reports in French using SIDRES.

In medical records, information organization follows a stable pattern and generally conforms to the SOAP model (such as age, sex, signs and symptoms, medical history, test results, etc). Introducing polarity in such typologies can lead to more fine-grained distinctions that have great informational value for clinical and research purposes. In fact, negation is a major source of poor precision in medical information retrieval systems (Rokach, Romano & Maimon, 2008; Averbuch, Karson, Ben-Ami, Maimon & Rokach, 2004).

Once we display a corpus of free-text clinical reports on SIDRES, we can create a set of twin categories : [1] <signs_symptoms> and [2] <signs_symptoms_n>, for encoding the presence and the absence of signs and symptoms, respectively ; [3] <history> and [4] <history_n>, for encoding the presence and the absence of medical history. We then incorporate the terms in the MeSH "C" descriptor to populate the SIDRES categories. While all four lists now feature the exact same content, introducing contexts on SIDRES can help solve this contradiction. We apply the French version of the NegEx patterns (Chapman, Bridewell, Hanbury, Cooper & Buchanan, 2001) for [2] (i.e., sans/pas de/absence de <signs of symptoms>). For [3] and [4], we introduce corpus-based syntactic patterns such as "with a history of" and "with no history of". [1] is left as the context-free, unmarked category.

3 Conclusion

We presented SIDRES, a new annotation tool that allows for the creation of categories and the introduction of contexts. We demonstrated the value of contexts for addressing negation in medical reports in French. Thanks to its flexibility, simplicity of use, and broad application possibilities, SIDRES can be easily incorporated to varied, small and large-scale annotation projects. SIDRES can be used for research purposes under a non-commercial license with an agreement of Avignon University. For commercial purposes, please contact Guillaume Gouvernet (guillaume.gouvernet@sattse.com).

References

AVERBUCH, M., KARSON, T., BEN-AMI, B., MAIMON, O. & ROKACH, L. (2004).Context-sensitive medical information retrieval. In Proc. of the 11th World Congress on Medical Informatics (MEDINFO-2004), pages 1–8. Citeseer. DOI : 10.3233/978-1-60750-949-3-282

CHAPMAN, W., BRIDEWELL, W., HANBURY, P., COOPER, G., BUCHANAN, B. (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries, Journal of Biomedical Informatics, Volume 34, Issue 5, pages 301-310, DOI: 10.1006/jbin.2001.1029.

ROKACH, L., ROMANO, R. & MAIMON, O (2008). Negation recognition in medical narrative reports. Inf Retrieval 11, 499–538. DOI : 10.1007/s10791-008-9061-0

TOMITA M. (1984). LR parsers for natural langages. *10th International Conference on Computational Linguistics*. *COLING*: 354-357. DOI : 10.3115/980491.980564