cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

agorantic Fédération de Recherche Culture, Patrimoines, Sociétés numériques
IRIT Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J
HCTI Héritages & Constructions dans le Texte & l'Image

TURN IT SIMPLE

SimpleText 2021

Simplification et Vulgarisation des Textes Scientifiques

Liana Ermakova, Eric San-Juan, Josiane Mothe

https://simpletext-madics.github.io/2021/

07 juillet 2021

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

- Les publications scientifiques sont difficiles à lire
- Lutter contre la désinformation
- Permettre de lire plus vite
- Faciliter l'accès aux
  - Non-natifs
  - Lecteurs plus jeunes
  - Citoyens avec des troubles de la lecture
- Améliorer des résultats des applications du TAL pour la pré-édition ou la traduction
- Pour:
  - communication scientifique
  - journalisme scientifique
  - communication politique
  - éducation

Motivation&Objectifs

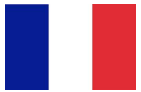
- Réunir une communauté scientifique interdisciplinaires
- Définition & Méthodes
- Contribuer à la réponse aux défis:
  - Sociétaux
  - Linguistiques
  - Techniques
  - D'évaluation
- Science ouverte et accessible

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Partenaires

- **Liana ERMAKOVA**
- **Eric SANJUAN**
- **Josiane MOTHE**

HCTI Héritages & Constructions dans le Texte & l'Image au XIXe siècle
agorantic Fédération de Recherche Culture, Patrimoines, Sociétés numériques
IRIT Institut de Recherche en Informatique de Toulouse CNRS - INP - UT3 - UT1 - UT2J

	SHS	Informatique
	• Élise Maturin • Michael Rinn • Radia Hannachi • Fidelia Ibekwe • Nicolas Poinsu • Helen McCombie HCTI École de Journalisme et de Communication d'Aix-Marseille Aix•Marseille Université BTU BUREAU DE TRADUCTION DE L'UBO	• Patrice Bellot • Sebastien Fournier • Diana Nurbakova • Ismail Badache • Stéphane Huet LIS Laboratoire d'Informatique & Systèmes LIRIS
	• Irina Ovchinnikova • Sílvia Lima Gonçalves Araújo CEHUM Centro de Estudos Humanísticos da Universidade do Minho Сеченовский Университет наук о жизни	• Jaap Kamps • Pavel Braslavski • John Rochford • Mike Unwalla Уральский федеральный университет имени первого Президента России Б.Н.Ельцина UNIVERSITY OF AMSTERDAM Techscribe® UMASS

TURN IT SIMPLE

cnrs GDR Groupement de recherche

MaDICS Masses de données, informations et connaissances en sciences

Données

ISTEX

Science

WIKIPEDIA
The Free Encyclopedia

dblp
computer science bibliography

The Guardian

easy text AI

yahoo! answers

LimeSurvey

ELI5

EasyCOVID-19
Making COVID-19 understandable for all

Dimensions

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Activités scientifiques

- SimpleText@CLEF - **21 - 24 Septembre 2021**
  - http://clef2021.clef-initiative.eu/index.php
  - 3 pilot tasks:
    i. Quelle information?
    ii. Quel contexte doit être donné?
    iii. Comment réduire la difficulté du texte sans altérer l'information?
- SimpleText@INFORSID - **1-4 juin 2021**
  - https://inforsid2021.sciencesconf.org/
  - Conférenciers invités:
    i. **Natalia Grabar**, https://pro.univ-lille.fr/natalia-grabar/
    ii. **John Rochford**, EasyText.AI and https://easycovid19.org/

- Évaluation
- Analyse des requêtes
- Définitions
- Stratégie
- Visualisation
- Données en français

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Analyse & création de données

- **Stratégie, source, évaluation:** Josiane Mothe, Eric SanJuan, Liana Ermakova + équipe CLEF
- **Analyse des requêtes:** Irina Ovchinnikova, Liana Ermakova, Diana Nurbakova
- **Définitions:** tous les partenaires
- **Préparation de données - Quelle information?** Eric Sanjuan, Nicolas Poinsu
- **Préparation de données - Quel contexte doit être donné?** Nicolas Poinsu, Ludivine Grégoire, Élise Mathurin, Irina Ovchinnikova, Sílvia Araújo
- **Préparation de données - Simplification?** Nicolas Poinsu, Liana Ermakova, Helen Mccombie
- **Visualisation du texte:** Radia Hannachi, Sílvia Araújo
- **Données en français:** Nicolas Poinsu

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Publications

- Actes des ATELIERS d' INFORSID - Dessinons ensemble le futur des systèmes d'information http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63
- Overview of SimpleText 2021 - CLEF Workshop on Text Simplification for Scientific Information Access
- What Science-Related Topics Need to Be Popularized? A Comparative Study

INFORSID

TURN IT SIMPLE

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

SimpleText@INFORSID'21

Rémi Cardon & **Natalia Grabar** - invited speaker

Recherche de phrases parallèles à partir de corpus comparables pour la simplification de textes médicaux en français

Mike Unwalla - industrial talk

Controlled language for text simplification: Concepts and implementation

Sílvia Araújo & **Radia Hannachi**

Pour une démarche de communication multimodale de données scientifiques : de la recherche documentaire à l'infographie via le mind mapping

Helen Mccombie

Could automatic text simplification assist correction-revision of scientific texts written by non-native English speakers?

John Rochford - invited speaker

Developing Simple Web Text for People with Intellectual Disabilities and to Train Artificial Intelligence

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Programme

9h - 9h30 : Ouverture

9h30 - 10h00 : **Alice Pennors** (Amstratgraph): "Comprendre la notion de "jumeau numérique": tentative de cartographie de la conception 2D/3D/4D"

10h00 - 10h30: **Helen MCCOMBIE** (UBO): "Simplifying meaning vs. deepening understanding: handling terminology in scientific English editing"

10h30 - 11h00: **Mathieu Rouault** (Grand Labo)

11h00 - 11h30: **Liana Ermakova** : SimpleText@CLEF overview

11h30 - 12h00: Table-ronde: "Évaluation de la difficulté du texte"

12h00 - 12h30: Présentation et lancement de l'atelier

TURN IT SIMPLE

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Merci !

Website: https://simpletext-madics.github.io/2021/

E-mail: simpletextworkshop@gmail.com

Twitter: https://twitter.com/SimpletextW

Google group: https://groups.google.com/g/simpletext

CLEF 2021 Workshop

SimpleText: Text Simplification for Scientific Information Access

Format & Call for Contributions

- Invited talks:
  - Xu Wei
  - John Rochford
  - Natalia Grabar
- Industrial talk: Mike Unwalla
- Presentations
- Interactive session

Types of contributions:

- **Participation in the pilot tasks!**
- Research & survey papers
- Position, discussion & demo paper
- Extended abstracts of published papers

Topics of interest (not exhaustive)

- Automated or computer-assisted scientific popularization/simplification
- Contextualization, search for background knowledge
- Terminology extraction
- Methods for assessing language complexity
- Methods for assessing information complexity
- Automatic summarization of scientific texts
- Daily digest generation
- Simplification of technical text, computer-assisted pre-editing
- Alteration and distortion of scientific information
- Automatic methods for scientific/data journalism

Pilot Tasks

Guidelines:

https://simpletext-madics.github.io/data/Guideline-SimplText-2021.pdf

TURN IT SIMPLE

CLEF 2021 BUCHAREST

Queries

- Press titles from *The Guardian* with manually extracted keywords
- Each keyword allows to extract at least 5 relevant abstracts
- Full text articles from The Guardian (link, folder query_related_content with full texts in the MD format)

Query 1: Digital assistants like Siri and Alexa entrench gender biases, says UN

https://www.theguardian.com/technology/2019/may/22/digital-voice-assistants-siri-alexa-gender-biases-unesco-says

Topic 1.1: Digital assistant

https://inex:qatc2011@guacamole.univ-avignon.fr/dblp1/_search?q="Digital assistant"&size=1000

Topic 1.2: Biases

https://inex:qatc2011@guacamole.univ-avignon.fr/dblp1/_search?q=biases&size=1000

TURN IT SIMPLE

Data

- Citation Network Dataset: DBLP+Citation, ACM Citation network (https://www.aminer.org/citation)
- DBLP full dump in the JSON.GZ format
- DBLP abstracts extracted for each topic in the following MD format

1551421219	2010	Online advertising has been fueling the rapid growth
2052650089	2012	Although online product reviews have emerged as an i
1571655962	2014	Purpose – The purpose of this paper is to investigat
2101571056	2005	This paper examines the practice of advertising with
2121552264	2012	The value proposition of mobile technology for educa

TURN IT SIMPLE

CLEF 2021 BUCHAREST

PILOT TASK 1 : Content Selection

Select passages to include in a simplified summary, given a query

Queries: titles of scientific journalism articles + keywords

Data: ElasticSearch index of Citation Network Dataset: DBLP+Citation, ACM Citation network

Evaluation: pooling, traditional IR metrics, unresolved anaphora,...

Potential problems:

- The information in a summary designed for an expert is different from those for the general audience
- Relevance of the source
- Unresolved anaphora
- ...

TURN IT SIMPLE

CLEF 2021 BUCHAREST

PILOT TASK 1 : Example

Input:

<topic>
<topic_id>1</topic_id>
<topic_text>Digital assistants like Siri and Alexa entrench ***gender biases****, says UN</topic_text>*
<keywords>
*<keyword>****Digital assistant*** *</keyword>*
*<keyword>****Biases****</keyword>*
</keywords>
</topic>

Expected output:

run_id manual topic_id doc_id passage rank

ST_1 1 1 3000234933 People are becoming increasingly comfortable using ***Digital Assistants*** *(DAs) to interact with services or connected objects. 1*

ST_1 1 1 3003409254 big data and machine learning (ML) algorithms can result in ***discriminatory*** *decisions against certain protected groups defined upon personal data like* ***gender****, race, sexual orientation etc. 2*

ST_1 1 1 3003409254 Such algorithms designed to discover patterns in big data might not only pick up any encoded societal ***biases*** *in the training data, but even worse, they might reinforce such* ***biases*** *resulting in more severe discrimination. 3*

TURN IT SIMPLE

CLEF 2021 BUCHAREST

PILOT TASK 2:
Searching for concepts to be explained

Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications,..)

Queries: titles of scientific journalism articles + keywords

Data: DBLP abstracts

Evaluation: NDCG?,...

Potential extension in future:

- Provide a context
- ...

TURN IT SIMPLE

CLEF 2021 BUCHAREST

PILOT TASK 2 : Example

Input:

<topic>
<topic_id>1</topic_id>
<topic_text>Digital assistants like Siri and Alexa entrench gender biases, says UN</topic_text>
<passage_id>1</passage_id>
<passage_text>Automated decision making based on big data and **machine learning** (**ML**) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc. Such algorithms designed to discover patterns in big data might not only pick up any encoded **societal biases** in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination.
</passage_text>
</topic>

Expected output:

Run_id	manual	topic_id	passage_id	term	rank
ST_1	1	1	1	machine learning	1
ST_1	1	1	1	societal biases	2
ST_1	1	1	1	ML	3

TURN IT SIMPLE

CLEF 2021 BUCHAREST

PILOT TASK 3: Language Simplification

Given a query, simplify passages from scientific abstracts

Queries: titles of scientific journalism articles + keywords

Data: DBLP abstracts

Evaluation: manual? Aggregated metrics?

Potential problems:

- Is it possible to simplify terminology? ⇒ Pilot task 2: background knowledge
- Out of scope of consideration: puns and idioms

TURN IT SIMPLE

CLEF 2021 BUCHAREST

PILOT TASK 3: Example

Input:

<topic>

<topic_id>1</topic_id>

<topic_text>Digital assistants like Siri and Alexa entrench gender biases, says UN</topic_text>

<passage_id>1</passage_id>

<passage_text>Automated decision making based on big data and machine learning (ML) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc. Such algorithms designed to discover patterns in big data might not only pick up any encoded societal biases in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination.

</passage_text>

</topic>

Expected output:

Run_id	manual	topic_id	passage_id	simplified_passage
ST_1	1	1	1	Automated decision-making may include sexist and racist biases and even reinforce them because their algorithms are based on the most prominent social representation in the dataset they use.

Organizers

- **Liana Ermakova,** HCTI - EA 4249, Université de Bretagne Occidentale (Brest, France)
- **Eric San-Juan,** Laboratoire d'Informatique d'Avignon,Institut de technologie d' Avignon (Avignon, France)
- **Josiane Mothe,** INSPE, Université de Toulouse, IRIT, UMR5505 CNRS (Toulouse, France)
- **Jaap Kamps,** Faculty of Humanities, University of Amsterdam (Amsterdam, Netherland)
- **Pavel Braslavski,** Combinatorial Algebra Lab, Ural Federal University, (Yekaterinburg, Russia)
- **Patrice Bellot,** Aix-Marseille Université - CNRS (LIS – INS2I) (Marseille, France)
- **Irina Ovchinnikova,** Institute of Linguistics and Intercultural Communication, Sechenov University (Moscow, Russia)
- **Diana Nurbakova,** LIRIS, Institut National des Sciences Appliquées de Lyon, (Lyon, France)

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Lancement de l'action SimpleText@MADICS'22

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Perspectives

- Définitions
- Jeux de données
  - + données en français
  - Informatique + médicine
- Metriques
- Atelier SimpleText@MADICS 2021 ⇒ Action SimpleText@MADICS 2022
- Workshop SimpleText@CLEF 2021 ⇒ Evaluation Lab SimpleText@CLEF 2022
- Atelier SimpleText@INFORSID
- Mots/Machines #4 https://motsmachines.github.io/2021/

Participez à SimpleText@CLEF !!!

TURN IT SIMPLE

cnrs GDR Groupement de recherche

MaDICS Masses de données, informations et connaissances en sciences

Données

ISTEX

Science

WIKIPEDIA
The Free Encyclopedia

dblp
computer science bibliography

The Guardian

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

yahoo! answers

arXiv.org

medRχiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

SCIENCES ET AVENIR

ELi5

Dimensions

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

SimpleText@CLEF'21 overview & lessons learned

- 43 registered teams
- 23 participants subscribed on our Google group
- 24 followers on Twitter
- Data was downloaded from the server by several participants, but no submitted runs → data can be reused
- We will enrich data prepared for the pilot tasks
- We continue to simplify passages
- New data can be released in early autumn 2021 → more time to potential participants
- New team members from humanities

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Partenaires/conférenciers potentiels

- Carole Chatelain, Sciences et Avenir
- William Audureau, Les décodeurs / Le Monde
- Núria Gala, projet ALECTOR (Aide à la LECTure pour améliORer l'accès aux documents pour enfants dyslexiques)
- Aymeric Poulain, Nereÿs
- Brest Métropole
- Florian Trigodet, MT180
- Ozvan Bocher, Science En Theizh

TURN IT SIMPLE

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Table ronde:
Évaluation de la difficulté du texte

TURN IT SIMPLE

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Questions

- Public général ? Différence entre les informations pour des experts et le public général ?
- Comment évaluer la difficulté des termes ? Quels termes doivent être expliqués ? Quels termes doivent être remplacés par les synonymes ?
- Quel type de définition il faut donner?
  - *Hydroxychloroquine (sulfate) is a medication used to prevent and treat malaria.*
  - *Hydroxychloroquine is an aminoquinoline that is chloroquine in which one of the N-ethyl groups is hydroxylated at position 2.*
  - *Hydroxychloroquine (sulfate) is a white, double-circle, film-coated, tablet imprinted with "PLAQUENIL".*
- Comment évaluer la difficulté du texte?
- Comment évaluer le niveau d'altération d'information acceptable?

TURN IT SIMPLE

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

Merci !

Website: https://simpletext-madics.github.io/2021/

E-mail: simpletextworkshop@gmail.com

Twitter: https://twitter.com/SimpletextW

Google group: https://groups.google.com/g/simpletext