CLEAR.TEXT Enhancing the Modernization Public Sector Organizations by Deploying Natural Language Processing to Make Their Digital Content CLEARER to Those with Cognitive Disabilities

Paloma Moreda[1], Beatriz Botella[1], Isabel Espinosa-Zaragoza[1], Elena Lloret[1], Tania J. Martin[1], Patricio Martínez[1], Armando Suárez[1] and Manuel Palomar[1]

[1]*GPLSI research group, Dept. of Software and Computing Systems*
Universidad de Alicante
CENID - Centro de Inteligencia Digital
03690, San Vicente, Alicante, España

Abstract
The CLEAR.TEXT project (TED2021-130707B-I00) researches how natural language processing technology can support the authoring of accessible content in Spanish for people with cognitive disabilities. Our main objective is to research, implement, deploy, evaluate, and ultimately provide robust technologies for natural language processing to support the authoring of accessible Spanish content for public sector organisations (at local, regional and national level) that is intelligible to people with cognitive disability, thereby widening their inclusion and empowerment in Europe. It is expected to impact positively the quality of life of people with cognitive disabilities, facilitating their access to educational, vocational, cultural, and social opportunities in public sector organisations

Keywords
Human Language Technologies, Natural Language Processing, Automatic Text Simplification, Automatic Summaries, Artificial Intelligence

1. Introduction

People with cognitive disabilities have significant limitations in their intellectual functioning and/or may also lack the ability to adapt to everyday situations. In fact, individuals with cognitive impairment have spoken and written word comprehension deficit that may include misinterpretation of literal meanings and difficulty understanding complex instructions. They are confused by idioms, figures of speech, abstractions, uncommon words, and lack of precision.

At present, natural language processing (NLP) technologies are mature enough to provide a sound basis for the development of components to automatically detect and remove obstacles to reading comprehension and generate additional content to facilitate reading comprehension. Hence, we start with the hypothesis that research, development and deployment of NLP technology can support the authoring of accessible content in Spanish for people with cognitive disabilities with a view to widening their inclusion and empowerment in Europe.

With this hypothesis in mind, the CLEAR.TEXT project[1] funded by the Spanish Government and the European Union and developed by the GPLSI research group[2] of the University of Alicante. The project focuses on research, implement, deploy, evaluate, and ultimately provide robust technologies for NLP to support the authoring of accessible Spanish content for public sector organisations (at local, regional and national level) that is intelligible to people with cognitive disability, thereby widening their inclusion and empowerment in Europe.

This, in turn, will improve the ability to access written information for all people, thereby reducing the risk of exclusion for those with cognitive disability. The project is expected to impact positively the quality of life of people with cognitive disabilities, facilitating their access to educational, vocational, cultural, and social opportunities in public sector organisations.

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations
moreda@dlsi.ua.es (P. Moreda); beatriz.botella@dlsi.ua.es (B. Botella); isabel.espinosa@ua.es (I. Espinosa-Zaragoza); elloret@dlsi.ua.es (E. Lloret); tania.martin@ua.es (T. J. Martin); patricio@dlsi.ua.es (P. Martínez); armando@dlsi.ua.es (A. Suárez); mpalomar@dlsi.ua.es (M. Palomar)

[1]https://cleartext.gplsi.es/
[2]https://cvnet.cpd.ua.es/curriculum-breve/grp/es/procesamiento-del-lenguaje-y-sistemas-de-informacion-(gplsi)/596

2. Objectives of the Project

The main objective of the CLEAR.TEXT project can be divided into the following specific objectives:

- O1. To analyse the main comprehension obstacles posed by the language used in the web content of Spanish public sector organizations, such as ministries and other government agencies, for people with cognitive disabilities.
- O2. To analysis of the needs of people with cognitive disabilities.
- O3. To research and adapt the COMPENDIUM System to the needs of public sector documentation.
- O4. To research, implement, deploy, and ultimately provide robust technologies to support the processing of structural complexity.
- O5. To research, implement, deploy, and ultimately provide robust technologies to support the processing of ambiguity in meaning.
- O6. To research, implement, deploy, and ultimately provide a robust text simplification system oriented towards public administration documentation.
- O7. To evaluate the simplification system.
- O8. To promote and disseminate the research results obtained from the project through different national and international media –including well indexed journals, conferences, seminars, etc., as well as exploit the potential for transferring this technology to society.

3. Human Resources

A multidisciplinary research team, comprising 5 specialists from the Computer Science and 3 from the Linguistics fields; in all 7 PhD holders and 1 technician. All researchers, 5 women and 3 men, are members of the GPLSI research group.

The team has extensive experience spanning 30 years of NLP-focused technological research and development, and specifically concerning the requested project in the areas of word sense disambiguation, anaphora resolution, coreference, named entity, lexical and syntactic analysis, text summarization, and text simplification -[?], [?], [?], [?], [?], [?], [?], [?].

4. Methodology and Work Plan

The research work that would be addressed in the CLEAR.TEXT project will have a duration of two years, starting from 1st September 2022.

The methodology has been organised into six main modules whose tasks and results are interlinked, as shown in Figure 1. An additional transversal seventh module focusses on coordination tasks.

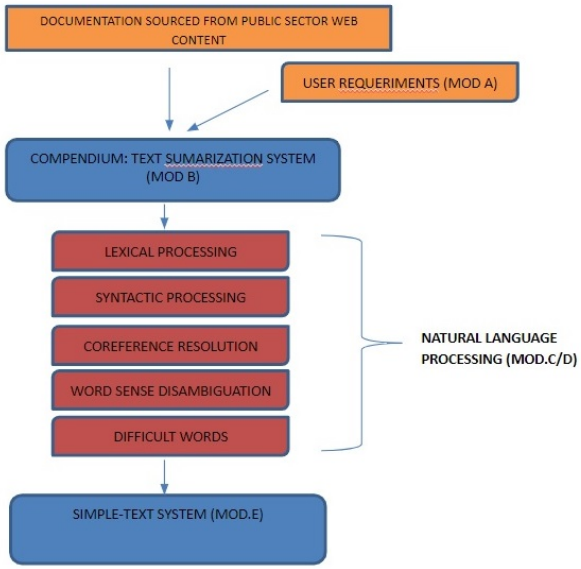

Figure 1: Connection between the proposed modules

Module A. Establishment and Analysis of User Requirements

This module is organized into two tasks which are detailed below.

Task A.1. Analysis of the main language phenomena. The objective of this task is to analyse the language phenomena, which will be referred to as obstacles to reading comprehension, both obstacles caused by ambiguity in meaning and obstacles caused by structural complexity, such us difficult words; morphologically, orthographically, and phonetically long or complex words; figurative language such as metaphor and idioms; long and syntactically complex sentences; or semantically highly specialised/technical words, among others.

Task A.2. User requirements. The objective of this task is to determine the users' requirements. For this, it is first necessary to determine and analyze the types of cognitive disabilities (dysphasia, Down syndrome, Fragile X syndrome and Autism Spectrum disorder) and, second, to establish user requirements and needs to facilitate information access for each type of cognitive disability.

Module B. Text summarization system: COMPENDIUM

Based on the expertise and previous research carried out by the team of the project, together with the advance-

ments of the state of the art in this task -[**?**], [**?**], [**?**]- the main goal of this module is to adapt the COMPENDIUM tool [**?**] to summarize administrative texts written in Spanish. Therefore, the following two tasks are included in this module.

Task B.1. Adapting COMPENDIUM to summarize documents of the public administrations. The main objective of this task is to adapt and fine-tune COMPENDIUM's approach to enable effective summarization of Spanish text documents generated by public sector organizations. The type of summary to be generated will be abstractive, which means that not only the relevant information will be extracted, but also, such relevant information will be paraphrased using different vocabulary and structures, guaranteeing that the same meaning in the generated text.

Task B2. Integrate COMPENDIUM in an accessible and easy-to-use platform. Once the COMPENDIUM tool has been adapted for Spanish public sector documents (task B.1), the goal of this task is to integrate the generated software into a user-friendly platform, so that it can be accessible and easy to use. For this, we will take into account the Web Accessibility Guidelines defined by the World Wide Web Consortium (W3C)[3], not only for the design of the Web interface, but also for the best way to output the content generated as the resulting summary, meeting the standards defined in: https://www.w3.org/WAI/standards-guidelines/wcag/.

Module C. Processing of structural complexity

This module aims at to analyse and investigate the state of the art of three main areas of natural language processing, and their application to the simplification of texts. These areas are lexical processing, syntactic processing, and discourse processing.

Task C.1. Lexical processing. The detection and removal of obstacles to reading comprehension caused by the individual words used in documents that users wish to access will build on a variety of related research. An exhaustive review of the state of the art will be carried out to determine the best lexical tools.

Task C.2. Syntactic processing. Many technologies to detect and remove syntactic obstacles to reading comprehension depend on syntactic parsing. Syntactic relations such as coordination and subordination, signaled by using conjunctions and punctuation, must be recognised accurately. An exhaustive review of the state of the art will be carried out to determine the best syntactic tools.

Module D. Processing of ambiguity in meaning

Various lines of research in the state of the art of processing ambiguity in meaning are relevant to the project. These include related work in three main areas: i)anaphora and coreference resolution, ii) the processing of word sense ambiguity, and iii) the processing of non-literal language.

Task D.1. Coreference resolution. One important process in the conversion of documents into a form facilitating reading comprehension is coreference resolution, in which links between pronouns and noun phrases referring to the same concept are automatically identified.

Task D.2. Word sense disambiguation. One prevalent source of ambiguity in language is word sense ambiguity, that is, when a single word may have any one of several meanings depending on the context in which it is used.

Task D.3. Processing of difficult words. The language used in public sector documents tends to be that of a sophisticated and technical nature. These documents usually contain archaisms, highly technical language, acronyms or aphorisms. The objective addressed by this task, via constructing or adapting a complex word processing tool, will be the detection and evaluation of the degree of complexity of these terms and their substitution by other equivalent words that are more easily understood.

Module E. Generation of simplex Documents

This project will be based on converting documents into an accessible personalised form for end-user profiles with cognitive difficulties.

Task E.1. Language processing components. One aspect of this functionality will be achieved by the implementation of language processing components aimed at removing obstacles to reading comprehension while preserving the meanings expressed by the original document. The language processing components implemented in the project will detect and remove obstacles caused by structural complexity (Module B) and ambiguity in meaning (Module C).

Task E.2. Generation of simplex documents. The second aspect of the conversion process will be the generation of content consist in textual data occurring in the document to supplement that textual data. This content will include document navigation tools, summaries, pre-reading questions, and images.

Task E.3. SIMPLE.TEXT System. Construction of a system that integrates the set of tools developed into a platform that is accessible and easy to use.

Module F. Evaluation and quality

Task F.1. Evaluation and quality. The objective of this task is to ensure the quality of the technology and results obtained in previous tasks. Both intrinsic and extrinsic evaluation methods will be performed. The implemented system (SIMPLE.TEXT system) will be evaluated with respect to the accessibility of the personalised documents that it generates by means of reading comprehension testing. The usability of the system by people with cognitive disabilities will be evaluated, and a qualitative research methodology employed to assess the contribution to inclusion resulting from the use of the technology.

Module G. Project coordination

[3] https://www.w3.org/WAI/fundamentals/accessibility-intro/

The objective of this module is to coordinate the flow of internal communication, evaluate progress and, if necessary, readjust the objectives. To successfully achieve this goal, one task is proposed (*Task G.1. Project monitoring*).

Module H. Project dissemination

This module is concerned with coordinating and supervising the dissemination of the project through publications and other means.

Task H.1. Dissemination of the project. The dissemination of the project, as well as the derived results, will be carried out during all stages of the project. The plan for communication, dissemination and exploitation of results will be articulated on three levels: i) information on the project and its development through a dedicated website; ii) scientific publications with research results, and iii) dissemination and exploitation of resources and prototypes created in the research project.

5. Scientific-Technical Impact

Language technologies are at the cornerstone of AI and are among those tools for which there will be the greatest demand in the next decade. Concerning the scientific and technical impact, our project focuses on research and development of technologies on NLP to support the authoring of accessible Spanish content for public sector organisations that is intelligible to people with cognitive disability. Applications derived from this research include the following: texts summarisation; texts simplification; lexical and syntactic analysis; anaphora resolution; word sense disambiguation; the creation of summarization reports providing explanations; and facilitating the comprehension of complex expressions (e.g. metaphors) in public sector organizations domains. The methods, models, resources, and systems that will be researched, developed and deployed in the project will attract the interest of the NLP and AI research communities.

6. Social and Economical Impact

We can identify the following positive social and economic impacts for people with cognitive disabilities that would result from being granted the opportunity to develop the proposed project:

- *Social impact* for those with cognitive disabilities:
  - Facilitates accessibility to digital information to promote social, and educational inclusion.
  - Narrows the digital divide by identifying barriers that prevent people with disabilities from accessing information on equal terms.
  - Promotes a cooperation between the technological and social fields, with a view to the design of technological solutions that consider the needs of people with disabilities.
  - Facilitates the daily actions of people with disabilities.
  - Widens inclusion and empowerment in Europe.
  - Improves the quality of life of those with cognitive disabilities, enhancing their access to educational, vocational, cultural, and social opportunities in Europe.
  - Promotes an independent life and the capability to realize personal goals.
  - Enables access to a meaningful education.
  - Encourages personal involvement in all decisions that affect their future.
  - Promotes participation in the benefits offered by cultural, recreational and sporting activities.
- *Economical impact* for those with cognitive disabilities:
  - Facilitates accessibility to digital information to promote economic and political inclusion.
  - Promotes full labor inclusion by 2024 through increased employment opportunities for individuals with cognitive disabilities and improved productivity by facilitating their ability to perform work-related tasks.
  - Widens inclusion and empowerment in Europe.
  - Enables participation in the services provided for promoting effective management of personal finances.
  - Provides equal access to and use of all facilities, services and activities in the public. sector organizations at local, regional and national level, such as filing tax returns, payment of fines, community charges, etc.

Aknowledgements

This research work is part of the CLEAR.TEXT project (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR.