The arText prototype: An automatic system for writing specialized texts

Iria da Cunha
Universidad Nacional de Educación a Distancia (UNED)
Senda del Rey 7, 28040
Madrid, Spain
iriad@flog.uned.es

M. Amor Montané
Universitat Pompeu Fabra
Roc Boronat 138, 08018
Barcelona, Spain
amor.montane@upf.edu

Luis Hysa
Universitat Pompeu Fabra
Roc Boronat 122, 08018
Barcelona, Spain
luis.hysa@gmail.com

Abstract

This article describes an automatic system for writing specialized texts in Spanish. The arText prototype is a free online text editor that includes different types of linguistic information. It is designed for a variety of end users and domains, including specialists and university students working in the fields of medicine and tourism, and laypersons writing to the public administration. ArText provides guidance on how to structure a text, prompts users to include all necessary contents in each section, and detects lexical and discourse problems in the text.

1 Introduction

In the field of Natural Language Processing (NLP), various types of linguistic information including phonological, morphological, lexical, syntactic, semantic and discourse-related features can be used to develop applications. To date, tools for writing texts have often been designed for general subject areas and included information on orthographic, grammatical and/or lexical aspects of the writing process. NLP researchers have tended not to study systems for structuring and writing specialized texts, although a few researchers have bucked this trend: Kinnunen et al. (2012) developed a system to identify and correct writing problems in English in several domains; Aluisio et al. (2001)'s system helps non-native speakers write scientific publications in English; the Writing Pal (Dai et al., 2011) and Estilector[1] systems help improve academic writing in English and Spanish, respectively; and LanguageTool[2] is an open source proofreading program for non-specialized texts in several languages. To our knowledge, none of the systems that are currently available have considered the specific characteristics of textual genres in specialized domains, such as medicine, tourism and the public administration.

Writing specialized texts is more challenging than writing general texts (Cabré, 1999). Textual, lexical and discourse features are an essential component of textual genres, such as medical research papers, travel blog posts, or claims submitted to the public administration. Against this backdrop, this article aims to present a prototype for an automatic system that provides assistance in writing specialized texts in Spanish. The arText system includes textual, lexical and discourse-related information, and is useful for different end users. It provides guidance on how to structure a text, prompts users to include all necessary contents in each section, and detects lexical and discourse problems in the text.

Da Cunha et al. (in press) determined the most frequent textual genres that pose the greatest writing challenges for three groups: specialists and university students in medicine and tourism, and laypersons writing to the public administration. ArText was designed to help these users write the 15 textual genres included in Table 1.

Section 2 describes the characteristics of the system and its modules. Section 3 explains how the system was evaluated, while Section 4 presents conclusions and future lines of research.

2 Description of the System

ArText is a free online text editor that anyone can use, with no registration required. The system was developed in a LINUX environment using an Apache server and a MySQL database. A variety of resources were utilized in the back end (BASH, PERL, and PHP, with a Laravel Frame-

[1] http://www.estilector.com/index.php.
[2] https://www.languagetool.org/.

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 57–60

Medicine	Research article Review article Medical history Abstract Bachelor's thesis
Tourism	Informative article Travel blog post Report Rules and regulations Business plan
Public Administration	Allegation Cover letter Letter of complaint Claim Application

Table 1: Specialized fields and textual genres included in arText.

work) and front end (HTML, CSS, JAVASCRIPT, with AJAX and JQUERY); Google Analytics is integrated into the site to measure traffic.

Documents can be exported in four formats: PDF, TXT, HTML and ARTEXT. Previously saved documents can be uploaded using the ARTEXT format, and the website includes a detailed user manual and a contact section for comments, questions and suggestions.

ArText can be accessed at `http://sistema-artext.com/`,[3] and has been optimized for the Google Chrome browser. To use arText, click on "Start using arText" and pick one of the 15 textual genres mentioned above. This brings you to the text editor, where you can start writing using the text editor and the three modules integrated into arText: Structure, Contents and Phraseology; Format and Spellchecking; and Lexical and Discourse-based Recommendations.

2.1 Module 1. Structure, Contents and Phraseology

The left-hand column helps users structure and draft documents. Its interactive template includes typical sections, contents and phraseology for each textual genre. This information was extracted from da Cunha and Montané (2016), a corpus-based analysis following van Dijk (1989)'s textual approach. Specifically, users can insert:

- Typical document sections
- Typical contents found in each section
- Phraseology related to each of these contents

The text editor displays the sections which typically appear in a given textual genre. For example, the template for a "claim" to be submitted to the public administration includes the following sections:

- Header
- Addressee
- Introductory clause
- Supporting details
- Request
- Closing

A drop-down menu in the left-hand column provides sample texts for each section, including section titles, where appropriate. For example, the "Supporting details" section includes two different contents:

- Grounds for the claim
- Attachments

When users click on a specific content, arText displays a list of sample phrases that can be incorporated into the final text. For example, "Attachments" includes the following phrases:

- Attached please find [document name].[4]
- The following supporting documents are attached: [list of documents].

Users can click on a stock phrase to include it in the text.

2.2 Module 2. Format and Spellchecking

The toolbar at the top of the screen includes an open source spellchecker (WebSpellChecker Ltd.) and various formatting options, e.g. to change font or font size; insert bullet points, images, tables and links; cut, copy and paste; print; and search. Since online storage is not provided, the user's manual includes instructions for uploading an image to Google Drive and inserting it into a document produced using arText.

2.3 Module 3. Lexical and Discourse-based Recommendations

By clicking on the review button in the right-hand column, users can see a series of lexical and discourse-related recommendations for improving their texts. These recommendations are derived from da Cunha and Montané (2016), which is based on Cabré (1999)'s Communicative Theory of Terminology and Mann and Thompson (1988)'s Rhetorical Structure Theory. The module includes a series of algorithms based on linguistic rules and two NLP tools: the Freeling shallow parser (At-

[3] A demo is available at `https://canal.uned.es/mmobj/index/id/54433`

[4] Users are instructed to fill in the information indicated in square brackets (e.g. names, dates and numbers).

serias et al., 2006) and the DiSeg discourse segmenter (da Cunha et al., 2012).

This module includes 11 main recommendations, all of which are displayed in the right-hand column, when appropriate. A subset of these recommendations is assigned to each textual genre, and all recommendations are adapted to the linguistic characteristics of each genre (da Cunha and Montané, 2016). Recommendations cover the following 11 topics:

1. Spelling out acronyms
2. Using acronyms systematically
3. Providing definitions
4. Using the passive voice
5. Using the 1st person systematically
6. Using subjectivity indicators
7. Repeating words
8. Using long sentences
9. Segmenting long sentences
10. Considering alternative discourse markers
11. Varying discourse markers

By clicking on a given recommendation, users can see a more detailed explanation and suggestions. In some cases, arText also highlights phrases or content in the text editor. For example, one lexical recommendation, "Spelling out acronyms," highlights acronyms that are not spelled out when they first appear in the text (i.e. arText would highlight "COPD" if "chronic obstructive pulmonary disease" did not appear next to this acronym the first time the term was used). Some recommendations also actively engage users in the revision process. For example, the recommendation "Repeating words" shows a list of repeated words; when users click on a word in the right-hand column, all occurrences of this word in the text are highlighted. This recommendation is not displayed for highly specialized textual genres (e.g. research articles and abstracts), since lexical variation is usually avoided in these types of texts (Cabré, 1999).

Some recommendations focus on the discourse level. For example, "Segmenting long sentences" highlights one or more long sentences; users can click to see suggestions for splitting them into shorter sentences. In this case, arText proposes these discourse segments in the right-hand column. The number of words used to determine long sentences differs for each textual genre, following da Cunha and Montané (2016).

Another discourse level recommendation refers to "Varying discourse markers." In this case, arText displays a list of discourse markers repeated in the text. When users click on one of these markers, all of its occurrences are highlighted, and a list of alternative discourse markers used to express the same relationship (e.g. Cause, Restatement, Contrast and Condition, etc.) is displayed in the right-hand column. For instance, for the discourse marker "that is," used to express Restatement, arText suggests the alternatives "in other words," "that is to say," "i.e." and "to put it another way."

3 Evaluation

Real and *ad hoc* texts were used to test arText's algorithms and linguistic rules and improve the system. Subsequently, the prototype was launched and data-driven and user-driven evaluations were conducted.

The data-driven evaluation was based on a test corpus with 24 texts corresponding to one textual genre from each domain; the corpus comprised eight medical abstracts, eight tourism-related informative articles and eight applications to the public administration. The linguistic characteristics of these texts were manually annotated, and the manual annotation and arText results were compared. Precision and recall were measured for a series of recommendations; the results are presented in Table 2.[5]

Recommendation ID	Precision	Recall
1	0.76	0.68
2	0.75	0.94
3	x	x
4	1	0.94
5: sing. verbal forms	0.87	0.97
5: pl. verbal forms	1	0.99
6	-	-
9	0.74	0.87
10	1	1

Table 2: Data-driven results.

Recommendation 7 did not apply in the medical subcorpus; in the tourism and public administration subcorpora, 91.67% and 94.70% of detected words, respectively, were repeated in the text. For Recommendation 8, 100% of highlighted sentences in the medicine and tourism subcorpora were long sentences according to the thresholds for abstracts and informative articles; no long sentences appeared in the public administration

[5] In light of the degree of specialization, Recommendation 4 did not apply for any of the textual genres included in the test corpus. No cases for which Recommendation 6 applies were found in the test corpus.

subcorpus, so this recommendation could not be tested for this genre. No cases of Recommendation 11 were found in the medical and administration subcorpora; in the tourism corpus, 100% of detected discourse markers were repeated in the text, and adequate alternatives were proposed.

The user-driven evaluation aimed to determine how useful arText is. A survey designed using Google Forms focused on accessibility, the usefulness of the three modules and general issues. Three doctors, three tourism professionals, and 25 laypersons completed the survey; all laypersons were between 30-50 years old and had both higher education experience and internet skills. In general, respondents found arText to be user-friendly and useful; 100% of them would recommend the system to other people. Respondents found the section on structure to be the most useful module, while the approach to uploading images was considered the system's greatest weakness.

4 Conclusions and Future Work

This paper describes a prototype of an automatic system to assist users in writing specialized texts. The online arText editor helps users draft texts for 15 textual genres in three specialized domains: medicine, tourism and the public administration. It lays out the structure for each section of the document, suggests appropriate contents and stock phrases for each section, and detects typical linguistic errors. This innovative system is the first tool that considers lexical, textual and discourse features for specific textual genres. Moreover, the arText project is based on the idea that academic research can be shared with and used constructively by the general public.

In the future, the results of the data-driven evaluation will be utilized to improve arText's algorithms. A second user-driven evaluation will include a broader population (e.g. students). Finally, arText may be adapted to other textual genres, specialized domains and languages.

Acknowledgments

This article is part of the "Automatic system to help in writing specialized texts in domains relevant to Spanish society" research project, funded by "2015 BBVA Foundation Grants for Researchers and Cultural Creators". It was also supported by a Ramón y Cajal contract (RYC-2014-16935). We would like to thank Josh Goldsmith for the proofreading of the text, and the evaluation survey respondents and the research groups ACTUALing and IULATERM for their support.

References

Sandra M. Aluísio, Iris Barcelos, Jandir Sampaio, and Osvaldo N. Oliveira Jr. 2001. How to learn the many unwritten "rules of the game" of the academic discourse: A hybrid approach based on critiques and cases to support scientific writing. *Proceedings of the IEEE International Conference on Advanced Learning Technologies, 2001*.

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. Freeling 1.3. syntactic and semantic services in an open-source nlp library. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 48–55.

M. Teresa Cabré. 1999. *La Terminología. Representación y comunicación*. Institut for Applied Linguistics, Barcelona.

Iria da Cunha and M. Amor Montané. 2016. Un análisis lingüístico de los géneros textuales del ámbito de la administración con repercusión en la vida de los ciudadanos. *Proceedings of the XV Symposium of the Ibero-American Terminology Network (RITerm 2016)*, pages 61–62.

Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. 2012. Diseg 1.0: The first system for spanish discourse segmentation. *Expert Systems with Applications*, 39(2):1671–1678.

Iria da Cunha, M. Amor Montané, and Alba Coll. in press. Detección de géneros textuales que presentan dificultades de redacción: un estudio en los ámbitos de la administración, la medicina y el turismo. *Proceedings of the 34th International Conference of the Spanish Association of Applied Linguistics*.

Jianmin Dai, Roxanne B. Raine, Rod Roscoe, Zhiqiang Cai, and Danielle S. McNamara. 2011. The writing-pal tutoring system: Development and design. *Journal of Engineering and Computer Innovations*, 2(1):1–11.

Tomi Kinnunen, Henri Leisma, Monika Machunik, Tuomo Kakkonen, and Jean-Luc Lebrun. 2012. Swan scientific writing assistant a tool for helping scholars to write reader-friendly manuscripts. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 20–24.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–288.

Teun A. van Dijk. 1989. *La ciencia del texto*. Paidós, Barcelona.