IRAZ: Easy-to-Read Content Generation via Automated Text Simplification

Thierry Etchegoyhen[1], Jesús Calleja Pérez[1] and David Ponce[1]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, 20009, Gipuzkoa, Spain

Abstract

Information complexity is a critical communication and integration barrier for large segments of the population. This situation is exacerbated by the ever increasing volumes of content generated in modern digital societies. The IRAZ project aims to develop a flexible solution for easy-to-read text generation, by means of automated text simplification, to support the production of accessible content. The project sets to produce new datasets in the field, created by professionals or via synthetic data generation. It also explores neural approaches to lexical, syntactic and end-to-end text simplification, offering configurable simplification hypotheses that can be post-edited by professional easy-to-read content creators. The project aims to develop a generic solution, with special emphasis on Basque and Spanish to provide further technological support for these two relatively under-resourced languages in the field.

Keywords

Accessibility, Easy-to-read, Text Simplification

1. Introduction

Digital transformation is rapidly impacting societies around the world, with ever increasing volumes of information being produced and shared in multiples languages and domains. Being able to understand this information is a requirement in modern society. Unfortunately, in many cases, the available information is not accessible to large portions of the population, due to its intrinsic complexity. Thus, a communication barrier exists for persons with cognitive dysfunction or impairment, older people, migrant populations and refugees, or people suffering from learning difficulties, among others.

Although estimates may vary depending on the country and the included segments of the population, it is typically estimated that more than 25% of the population faces reading and comprehension difficulties in a country like Spain [1]. For such a large portion of the population, the inability to properly access the information in essential domains such as health, education, culture or media, can result in social exclusion in many sectors and activities.

The easy-to-read method[1] aims to facilitate information access by presenting it under a series of guidelines, including: favour the use of simple words and short sentences with less than 15 words, employ direct language, and present each sentence separately, with one grammatical segment per line.[2] This method is one of the main tools to facilitate the communication of information across domains. However, current processes are mainly performed manually, without technological support for the most part. This state of affairs hinders the production of new accessible content and the development of a more inclusive society.

Project IRAZ[3] aims to address the current limitations in easy-to-read content production, via the development of an integrated solution based on automated text simplification (ATS) technology. Its main objective is thus the development of an application that will assist easy-to-read content creators, by providing simplified versions of texts and allowing experts to post-edit system suggestions. The collected post-edited data will further allow the training of improved versions of the text simplification models that will be developed within the project.

2. Consortium and Funding Body

IRAZ is partially funded by the Basque Government via the Hazitek 2022 program of the Spri Group, as an industrial research project under Grant Agreement ZL-2022/00788. The project started in April 2022 and will finalise in December 2024.

The consortium includes the following participants: Vicomtech[4] as the research centre leading research and development activities; Merkatu Digital[5], as project co-

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations

tetchegoyhen@vicomtech.org (T. Etchegoyhen); jcalleja@vicomtech.org (J. C. Pérez); adponce@vicomtech.org (D. Ponce)

0009-0008-9642-5658 (J. C. Pérez)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

[1]Also known as *Easy Read*; *Lectura Fácil* in Spanish.

[2]https://www.inclusion-europe.eu/easy-to-read-standards-guidelines

[3]The name stems from *irakurketa erraza*, which means easy-to-read in Basque.

[4]https://www.vicomtech.org

[5]https://www.merkatu.com

CEUR-WS.org/Vol-3516/paper13.pdf

ordinator; Gureak Marketing[6]; Lantegi Batuak[7]; Lectura Fácil Euskadi[8]; and Merkatu Interactiva[5]. It is worth noting that the project includes leading experts in easy-to-read content creation and dissemination in the Basque Country.

3. State of the Art

As noted above, most easy-to-read content is typically created in a manual fashion by experts in the field, without relevant technological support. A limited number of projects, e.g., the Easy Reading project[9], have attempted to include this type of support for easy-to-read content generation and dissemination. Easy-to-read content usually involves adaptation beyond text simplification, in particular via the provision of additional information, such as simple explanations of complex concepts. Nonetheless, automated text simplification, defined as the reduction of complexity of a given text while retaining the original content and information [2], is typically viewed as a key enabling technology to facilitate access to textual information for people with reading difficulties. Projects such as PSET [3] and Simplext [4], or, more recently, ConMuTeS[10] and SimpleText[11], have set to provide this type of support via ATS technology.

Texts may be simplified at different levels: lexical, replacing complex words with simpler synonyms; syntactic, transforming complex sentences involving coordination or different types of modifiers into separate simple sentences, or transforming passive voice into active, among other operations; and conceptual, tackling coreference resolution, for example. Earlier approaches attempted to model these transformations via computational rules, hand-crafted [5] or inferred from aligned corpora [6]. Later data-driven techniques such as Statistical Machine Translation led to formulating the simplification problem as an end-to-end monolingual translation task, using corpora of aligned complex and simple sentences [7].

As in most natural language processing fields, data-driven approaches based on artificial neural networks and deep learning have become the dominant paradigm in ATS research, in recent years. For lexical simplification (LS), for example, LS-BERT [8] is currently a standard baseline, based on a BiLSTM model for complex word identification (CWI) and a pretrained BERT model [9] for substitute candidate generation. For Spanish, one of the languages of interest in IRAZ, Alarcón et al. [10] use contextual vectors extracted from pretrained mBERT and BETO models, among other features, to perform SVM-based CWI and use these language models (LMs) for substitution generation and ranking. In recent results from the TSAR-2022 Shared Task [11], participants have mostly used neural LMs for the LS task. The best results on this task, for English, were obtained via prompts fed to the very large generative pretrained language model GPT-3 [12]. Current LS approaches mainly use task-specific corpora for system benchmarking [13, 14].

Most current ATS models beyond LS employ end-to-end (E2E) architectures, which attempt to model a complete transformation from complex to simple text. Different architectures have been explored along these lines, e.g., LSTMs [15] or Semantic Encoders Vu et al. [16], although, in recent years, Transformer models [17] have become ubiquitous for ATS as well. Thus, Zhao et al. [18] combine Transformers and paraphrase rules for the task, whereas Martin et al. [19] use parameter tokens to control the output sequence in terms of character length ratio and Levenshtein distance in their ACCESS model. Variants of the latter approach have been proposed with pretrained BART models [20] and fine-tuned T5 models [21, 22]. Alternatively, Omelianchuk et al. [23] use a RoBERTA model to tag the input sequence with keep, delete or append tokens, among others, moving from a generative task into a sequence labelling task.

An important limitation for ATS modelling is the scarcity of training corpora, particularly for E2E models. Most datasets are available only for English and the largest are derived from alignments of Wikipedia and Simple Wikipedia content [24, 25]. Xu et al. [26] highlight critical issues with Simple Wikipedia for ATS and introduced Newsela, a professionally-produced corpus for English and Spanish. For Basque, only two small hand-crafted datasets are currently available, one in the science popularisation domain [27], the other on news from the Irekia Open Government portal[12] [28].

To overcome data scarcity, different approaches have been recently proposed for ATS. Thus, Surya et al. [29] described an unsupervised method based on a shared encoder and two attentional-decoders with discrimination-based losses and denoising, which can be trained on unlabeled text data or in a weakly supervised fashion with a few aligned examples. The creation of synthetic datasets is another alternative to address training data scarcity. Along these lines, Lu et al. [30] exploit machine translationese to build pseudo-parallel datasets for the task. In Kim et al. [31], existing parallel datasets are exploited to create one-to-many monolingual parallel corpora, via machine translation, allowing the training of models that learn to split complex sentences into simpler ones.

[6] https://www.gureakmarketing.com
[7] https://www.lantegibatuak.eus
[8] https://lecturafacileuskadi.net
[9] https://www.easyreading.eu/the-project/
[10] https://www.upf.edu/web/conmutes
[11] https://anr.fr/Project-ANR-22-CE23-0019
[12] https://www.irekia.euskadi.eus/en

4. IRAZ

4.1. Objectives and Challenges

The main objective of project IRAZ is the development of a software solution to support the generation of easy-to-read textual content, by means of automated text simplification.

Although the goal is to develop a generic solution which could support any language, in principle, a secondary objective of the project is the development of relevant resources for two specific languages, which are critical to the participants of the project: Spanish, for which limited resources are currently available, and Basque, for which there exist only two small datasets, as previously indicated.

As described in the previous section, ATS is a challenging field in many respects. First, there is a significant lack of resources across languages and domains, which hinders the development of data-driven models on a par with those obtained in neighbouring fields such as neural language modelling and machine translation.

4.2. Approach

Given the current challenges and limitations in the field of ATS, the project adopted a multi-pronged approach for the development of a flexible solution which could both (i) benefit easy-to-read content creators, and (ii) help advance the state-of-the-art in the field. The main components of our approach are summarised below.

4.2.1. Data creation

The project involves specific resource collection and generation activities. Documents consisting of complex or simplified data[13] will thus be collected and processed within various cycles of the project. This activity will lead to the generation of new datasets in different domains, in the two selected languages. It includes both the collection of proprietary data from project participants' repositories, for which exploitation right within the project have been cleared, and the preparation of new datasets from public sources, such as Irekia (*op. cit.*) for Basque and Spanish news.

Considering the current lack of resources across the board to train end-to-end neural simplification models, part of the project activities are focused on synthetic data generation. We thus investigate different methods to generate synthetic datasets of complex-simple pairs, via synthetic data merger, paraphrase generation or artificial one-to-many data generation along the lines of [31].

All data are automatically extracted and aligned, using standard tools such as CATS [32] or in-house processing scripts developed within the project.

4.2.2. Simplification Methods

Instead of providing one-size-fits-all simplification, irrespective of a given adaptation task, IRAZ aims to investigate different types of ATS methods separately, and allow users to select and test different ATS models individually or in combination, depending on the content at hand. For example, with specific types of content, performing only lexical simplification might be an optimal choice, if other types of transformation do not provide accurate results, or if the user is only looking for this type of simplification. Alternatively, for other types of content, end-to-end ATS models may provide sufficient quality and be used directly to generate simplification hypotheses.

A significant part of the research activities within the project will thus address the following three main types of ATS methods:

- *Lexical simplification*: A specific set of models and methods will be developed for multilingual lexical simplification, based on statistical methods, embeddings and pre-trained neural language models such as BERT or XLM-R [33]. We explore in particular the impact of cascading lexical replacements, whole-word vs. subword masking strategies, for agglutinative languages like Basque in particular, and contextual coherence of lexical substitution.
- *Syntactic simplification*: Under this designation, we include all methods and models that specifically tackle the transformation of complex sentences into separate simple sentences. This includes the extraction of complex modifiers, such as relative or appositive clauses, and splitting coordinated or juxtaposed sentences, for instance. The project will focus on neural models for this task, with a strong emphasis on the creation of artificial datasets from which syntactic simplification may be modelled.
- *End-to-end simplification*: Under this banner are all neural text simplification models that perform the task in and end-to-end fashion. This includes the development and evaluation of models based on control parameter tokens, synthetic corpora, or pre-trained language models fine-tuned for ATS tasks. Note that the data generated via the first two approaches above are also exploited for synthetic data generation to train the end-to-end ATS models.

[13] For convenience, we refer to both simplified and adapted content as *simplified* content, although adaptation can diverge from strict simplification (which, by definition, should not alter the information in the original text), via additional explanations or complex concept removal, for instance.

The IRAZ application has been designed to support flexible on-the-fly access to different type of ATS models, in isolation or in combination via model ensembling. In addition to models centred on specific aspects of text simplification, it is worth noting that the project also addresses automated text segmentation methods to support easy-to-read text adaptation.

4.2.3. Post-editing

Even more so than in companion fields such as machine translation, where high quality may be achieved from existing parallel training resources, ATS output requires post-editing by experts prior to its publication as easy-to-read content, for two main reasons. First, the overall quality achieved by state-of-the-art models, in particular for languages like Spanish or Basque, is still too low for most results to be directly exploited in professional settings. Secondly, as previously noted, easy-to-read content may require further adaptation of the simplification suggestions, be they at the lexical or grammatical level.

The project takes these current technological limitations into account, by providing a Web-based user interface, where users can post-edit the suggested simplifications generated with the underlying ATS models. Each sentence in the original text is automatically split and its automatic simplification presented in a separate editable window, the two being presented side by side. Users can freely post-edit the simplified text and download the complete text once the overall editing process is completed. The professionally post-edited data will then be collected to re-train or fine-tune ATS models.

4.3. Initial Results

The first phase of the project, in 2022, centred on the design of the solution, based on collected user requirements, an initial data collection phase, and preliminary research on core ATS tools and methods, with specific emphasis on the main use-cases of the project for the Basque and Spanish languages.

In its second phase, which debuted in 2023, the project has entered the core research and development cycles defined in the work plan, where ATS modelling and prototype development is taking place, along with further data collection and preparation.

The initial results obtained to date are summarised below:

- New corpora for Basque and Spanish, based on data collected from the Irekia portal. The data were collected from Web crawls, with automatic pairing of complex and corresponding simplified documents, text extraction and many-to-many sentence alignment.[14] We also prepared initial datasets from project participants' repositories, with further data to be collected in the remaining phases of the project.
- Initial models for lexical simplification and comparative results for Basque and Spanish, based on both LS-BERT and our own BERT-based variant which includes additional features and candidate substitution methods.
- Initial models for one-to-many sentence transformation, following the BISECT approach adapted to Basque and Spanish via high-quality machine translation of the original English datasets and data selection.
- Initial models for end-to-end neural text simplification, based on Transformer encoder-decoder architectures and training on mixtures of monolingual and simplified data.
- Initial automated text segmentation methods, based on pretrained mBERT and predictions of chunk completion via token masking.
- Initial development of the core components, API and UI of the IRAZ solution.

Current initial results provide a solid basis for the remaining planned research and development activities of the project.

5. Conclusions

We have described the IRAZ project, whose main goal is the development of a flexible solution to support easy-to-read content generation, by means of automated text simplification. The project targets professional easy-to-read content creators, who lack technological support in actual practice. The project aims to help optimise current content generation processes in the field, thus enhancing the creation of accessible content for the large segments of the population in critical need of this type of content. Research and development goals and activities have been defined in close collaboration with the professional end-users in the field who actively participate in the project.

Acknowledgments

IRAZ is partially funded by the Basque Business Development Agency, SPRI, under Grant Agreement ZL-2022/00788. We wish to thanks all participants of the project for their contributions and insights. We also wish to thank the anonymous SEPLN reviewer for their comments and suggestions.

[14] We aim to share the prepared datasets with the research community in the near future.

References

[1] S. Štajner, Automatic text simplification for social good: progress and challenges, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (2021) 2637–2652.

[2] A. Siddharthan, A survey of research on text simplification, ITL-International Journal of Applied Linguistics 165 (2014) 259–298.

[3] J. Carroll, G. Minnen, Y. Canning, S. Devlin, J. Tait, Practical simplification of english newspaper text to assist aphasic readers, in: Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, Association for the Advancement of Artificial Intelligence, 1998, pp. 7–10.

[4] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making it Simplext: Implementation and evaluation of a text simplification system for Spanish, ACM Transactions on Accessible Computing (TACCESS) 6 (2015) 1–36.

[5] R. Chandrasekar, C. Doran, S. Bangalore, Motivations and methods for text simplification, in: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, 1996, pp. 1041–1044.

[6] R. Chandrasekar, B. Srinivas, Automatic induction of rules for text simplification, Knowledge-Based Systems 10 (1997) 183–190.

[7] L. Specia, Translating from complex to simplified sentences, in: Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings 9, Springer, 2010, pp. 30–39.

[8] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, X. Wu, Lexical simplification with pretrained encoders, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8649–8656.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[10] R. Alarcón, L. Moreno, P. Martínez, Exploration of Spanish word embeddings for lexical simplification., in: CTTS@ SEPLN, 2021, pp. 29–41.

[11] H. Saggion, S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, M. Zampieri, Findings of the TSAR-2022 shared task on multilingual lexical simplification, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual), 2022, pp. 271–283.

[12] D. Aumiller, M. Gertz, UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification?, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual), 2022, pp. 251–258.

[13] D. Ferrés, H. Saggion, ALEXSIS: a dataset for lexical simplification in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3582–3594.

[14] S. Uchida, S. Takada, Y. Arase, CEFR-based lexical simplification dataset, in: Proceedings of International Conference on Language Resources and Evaluation, volume 11, European Language Resources Association, 2018, pp. 3254–3258.

[15] X. Zhang, M. Lapata, Sentence simplification with deep reinforcement learning, in: EMNLP 2017: Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 584–594.

[16] T. Vu, B. Hu, T. Munkhdalai, H. Yu, Sentence simplification with memory-augmented neural networks, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 79–85.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[18] S. Zhao, R. Meng, D. He, A. Saptono, B. Parmanto, Integrating transformer and paraphrase rules for sentence simplification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3164–3173.

[19] L. Martin, É. V. De La Clergerie, B. Sagot, A. Bordes, Controllable sentence simplification, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4689–4698.

[20] L. Martin, A. Fan, É. V. De La Clergerie, A. Bordes, B. Sagot, Muss: Multilingual unsupervised sentence simplification by mining paraphrases, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1651–1664.

[21] K. C. Sheang, H. Saggion, Controllable sentence simplification with a unified text-to-text transfer transformer, in: Proceedings of the 14th International Conference on Natural Language Generation,

2021, pp. 341–352.
[22] S. Štajner, K. C. Sheang, H. Saggion, Sentence simplification capabilities of transfer-based models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 12172–12180.
[23] K. Omelianchuk, V. Raheja, O. Skurzhanskyi, Text simplification by tagging, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 2021, pp. 11–25.
[24] Z. Zhu, D. Bernhard, I. Gurevych, A monolingual tree-based translation model for sentence simplification, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010, pp. 1353–1361.
[25] W. Coster, D. Kauchak, Simple english wikipedia: a new text simplification task, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 665–669.
[26] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297.
[27] I. Gonzalez-Dios, M. J. Aranzabe, A. Díaz de Ilarraza, The corpus of Basque simplified texts (CBST), Language Resources and Evaluation 52 (2018) 217–247.
[28] I. Gonzalez-Dios, I. Gutiérrez-Fandiño, O. M. Cumbicus-Pineda, A. Soroa, IrekiaLFes: a new open benchmark and baseline systems for Spanish automatic text simplification, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), 2022, pp. 86–97.
[29] S. Surya, A. Mishra, A. Laha, P. Jain, K. Sankaranarayanan, Unsupervised neural text simplification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2058–2068.
[30] X. Lu, J. Qiang, Y. Li, Y. Yuan, Y. Zhu, An unsupervised method for building sentence simplification corpora in multiple languages, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 227–237.
[31] J. Kim, M. Maddela, R. Kriz, W. Xu, C. Callison-Burch, BiSECT: Learning to split and rephrase sentences with bitexts, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6193–6209.
[32] S. Stajner, M. Franco-Salvador, S. P. Ponzetto, P. Rosso, H. Stuckenschmidt, Sentence alignment methods for improving text simplification systems, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers, 2017, pp. 97–102.
[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451.