

# SciNews: From Scholarly Complexities to Public Narratives – A Dataset for Scientific News Report Generation

Dongqi Liu, Yifan Wang, Jia Loy, Vera Demberg

Department of Computer Science

Department of Language Science and Technology

Saarland Informatics Campus, Saarland University, Germany

{dongqi, yifwang, jialoy, vera}@lst.uni-saarland.de

## Abstract

Scientific news reports serve as a bridge, adeptly translating complex research articles into reports that resonate with the broader public. The automated generation of such narratives enhances the accessibility of scholarly insights. In this paper, we present a new corpus to facilitate this paradigm development. Our corpus comprises a parallel compilation of academic publications and their corresponding scientific news reports across nine disciplines. To demonstrate the utility and reliability of our dataset, we conduct an extensive analysis, highlighting the divergences in readability and brevity between scientific news narratives and academic manuscripts. We benchmark our dataset employing state-of-the-art text generation models. The evaluation process involves both automatic and human evaluation, which lays the groundwork for future explorations into the automated generation of scientific news reports. The dataset and code related to this work are available at <https://dongqi.me/projects/SciNews>.

**Keywords:** Scientific News Report Generation, Natural Language Generation, Text Summarization

## 1. Introduction

**Why Studying Scientific News Report Generation is Valuable:** Scientific publications capture the latest advancements and discoveries in the realm of science, but often necessitate a significant level of academic background, posing obstacles for the general public without specialized knowledge (Saikh et al., 2020; Wright and Augenstein, 2021; August et al., 2022; Wright et al., 2022a). In a bid to bridge this knowledge gap, science journalists are endeavoring to translate intricate scientific nuances and breakthroughs into concise and accessible language (Polman and Hope, 2014; Majetic and Pellegrino, 2014; Li et al., 2017; Hoque et al., 2022). This initiative seeks to promote a profound engagement between the public audience and scientific literature (Ravenscroft et al., 2018; Vadapalli et al., 2018; August et al., 2020). Figure 1 illustrates how scientific news reports/narratives may help to increase the accessibility of scientific discoveries by using simplified language, examples, and explanations for technical terms (e.g., “cybersickness” / “feeling nauseous or disorientated”). Regrettably, the pursuit of automated generation of scientific news reports faces challenges due to the insufficient availability of parallel corpora. Thus, this paper proposes (i) a new task, Automated Scientific News Report Generation (SNG), and (ii) a novel dataset, SciNews, designed for this task.

**Similarities and Differences with Text Summarization and Text Simplification:** Text summarization emphasizes the reduction of textual volume whilst preserving main information, without altering linguistic complexity (Liu et al., 2023b; Pu

Figure 1: An example of an academic paper paired with its news report.

et al., 2023; Hosking et al., 2023; Cho et al., 2022; Goyal et al., 2022; Pu et al., 2022; Lai et al., 2022; See et al., 2017), while text simplification focuses on employing simplified lexicon and syntax to enhance readability (Pu and Demberg, 2023; Nisioi et al., 2017; Sulem et al., 2018; Blinova et al., 2023; Laban et al., 2023; Garimella et al., 2022; Crippwell et al., 2022). The SNG intertwines these processes, requiring both the simplification of complex concepts to more comprehensible forms and the extraction of pivotal insights from source materials (Alambo et al., 2020; August et al., 2020; Dong et al., 2021; August et al., 2022; Tan et al., 2023).

Unlike previous efforts mainly focusing on the































