Searching for explanation of difficult scientific terms

Majda ENNACIRI

Abstract

Understanding scientific texts is an essential skill for successful learning in school and throughout life (project...). However, non-experts (scientists who are interested in scientific documents from disciplines in which they are not experts) encounter significant difficulties in understanding them. This is due to key words that are difficult to understand without prior knowledge, linguistic complexity, structure and length of scientific articles. In this work, our goal is to generate an explanation for a given difficult scientific term in order to help a user understand the text (definitions, examples, use cases,...). First, we trained an AI model to predict a context to a given term. Then, we compared these results with different baselines.

Keywords

text simplification, reading comprehension, automatic natural language processing, information search, simplified text, difficult scientific terms,

1. Introduction

The lack of basic knowledge can become an obstacle to reading comprehension and reduce access to information. Simplification of scientific texts can then appear as an aid because its objective is to make complex content more easily understandable by establishing links with the basic lexicon. Traditional methods of simplifying texts can eliminate complex concepts and constructions. Furthermore, simplification is about reducing the complexity of a text while retaining the original information and meaning.

As part of the CLEF 2022 SimpleText lab[1] competition, which aims to simplify scientific texts, we solved task 2, which is the search for difficult words that make it difficult to understand the texts. Despite much research, these terms remain a difficult task to define. In general, these terms can be defined as the concepts of a domain. But, such a definition leaves room for several questions about the nature of the terms, the problem lies in practical aspects such as the length of the terms and theoretical considerations about the difference between words and terms. This leads to many problems, from data collection to extraction and evaluation.

First, among the methods that do term extraction, there are linguistic methods that rely on linguistic information such as POS models and chunking. In addition, statistical methods, which use frequencies compared to a reference corpus, and hybrid methods, which combine the two methods just mentioned. They tend to be better in terms of performance compared to other methods.These methods select candidate terms based on their POS model and rank them using statistical metrics. Secondly, the advancement of machine learning techniques has made it more

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

ennacirimajda@gmail.com (M. ENNACIRI)

CEUR Workshop Proceedings (CEUR-WS.org)

complicated to classify such simple methodologies mentioned before[2]. Finally, Jurassic-1 [3] language models, with 178B parameters for J1-Jumbo capable of transforming existing text, e.g., in the case of summarization or prediction of difficult words.

The following sections begin with a brief overview of the methodology used (IDF, PMI) to perform difficult word extraction and the datasets used. In addition, the next section contains the results obtained. The last section is devoted to a discussion and conclusions.

2. Methods

In order to determine the terms that require explanation and contextualization to help a reader understand a complex scientific text, for example, in relation to a query, the terms that need to be contextualized (with a definition, an example and/or a use case). To do this, we computed the IDF score (which gives us the frequency of use of a word and from which we ranked these words by order of difficulty) and the PMI scoreon on a dataset that we will present next.

After computing the scores, we set a threshold from which we extracted the difficult words. That is, words that have an IDF score greater than or equal to this threshold (we set this score from the training set) are considered difficult words. Moreover, we have two types of difficult words: term or sentence, in order to extract the complex sentences, we computed the PMI of each bigram and extracted those with the highest PMI in a given content.

We classified them into 3 (difficulty scores are: 1, 2 and 3) and 5 (difficulty scores are: 1, 2, 3, 4 and 5). In addition, the sentences whose scores do not exceed the thresholds are considered easy to understand.

2.1. IDF (Inverse Document Frequency) score

Inverse Document Frequency (IDF) is a measure of how often a word is used, i.e. how much information the word provides. The higher its score, the more important the word is. It is the inverse fraction of documents that contain the word (obtained by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that quotient).
IDF of a term t is computed as:

$$IDF(t) = \ln(\frac{N}{N_t})$$

where N is the total number of documents in the corpus, and N_t is the number of documents containing the term t. [4]

2.2. The Pointwise Mutual Information (PMI) Criterion

Pointwise Mutual Information (PMI) has proven to be a useful association measure in many natural language processing applications, such as collocation extraction and word space models.

The idea of PMI is to quantify the probability of co-occurrence of two words. The algorithm computes the (logarithmic) probability of co-occurrence,divided by the product of the single occurrence probability, as follows:

$$PMI(a,b) = \log(\frac{P(a,b)}{P(a)P(b)})$$

With a and b are the terms that we want to calculate their PMI. knowing that, when ’a’ and ’b’ are independent, their joint probability is equal to the product of their marginal probabilities, when the ratio is equal to 1 (so the logarithm is equal to 0), this means that the two words together do not form a unique concept: they co-occur by chance. [5]

2.3. Dataset

2.3.1. Train Dataset

The data are extracted under the two topics: Medicine and Computer Science, as these two areas are the most popular on the forums. As in 2021, for computer science, they use the scientific abstracts from the Citation Network dataset: DBLP+Citation, ACM Citation network. A student who is proficient in technical writing and translation manually annotated each sentence extracting difficult words.[6]

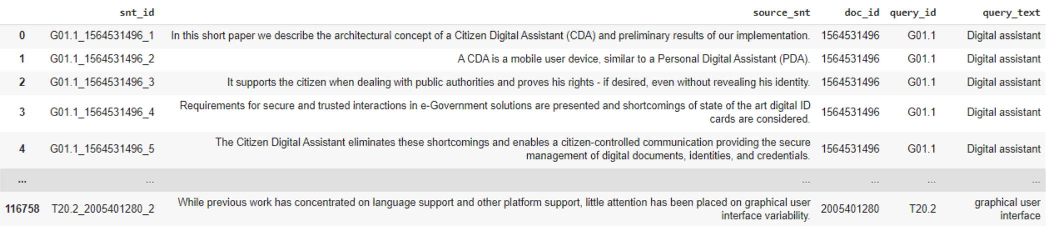
2.3.2. Test Dataset

To construct the test data, 116,763 sentences were extracted from DBLP summaries with the following queries:

- **Input and output formats.** The input for the train and the test data was provided in JSON and CSV formats with the following fields:
- **snt id** a unique passage (sentence) identifier.
- **source snt** passage text
- **doc id** a unique source document identifier.
- **query id** a query ID.
- **query text** difficult terms should be extracted from sentences with regard to this query.

[6]

Input examples (CSV format):

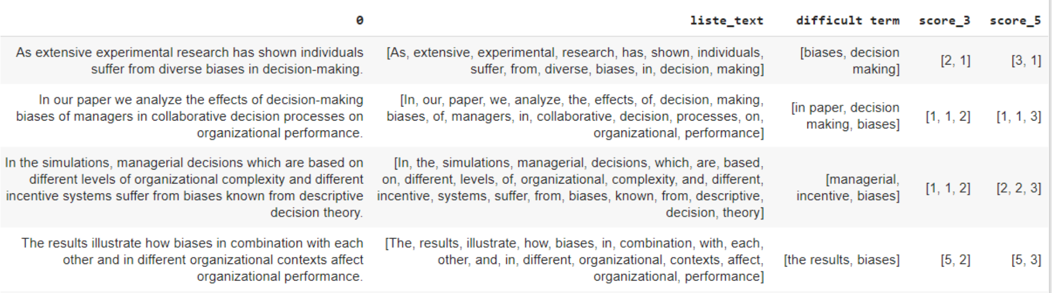


	snt_id	source_snt	doc_id	query_id	query_text
0	G01.1_1564531496_1	In this short paper we describe the architectural concept of a Citizen Digital Assistant (CDA) and preliminary results of our implementation.	1564531496	G01.1	Digital assistant
1	G01.1_1564531496_2	A CDA is a mobile user device, similar to a Personal Digital Assistant (PDA).	1564531496	G01.1	Digital assistant
2	G01.1_1564531496_3	It supports the citizen when dealing with public authorities and proves his rights - if desired, even without revealing his identity.	1564531496	G01.1	Digital assistant
3	G01.1_1564531496_4	Requirements for secure and trusted interactions in e-Government solutions are presented and shortcomings of state of the art digital ID cards are considered.	1564531496	G01.1	Digital assistant
4	G01.1_1564531496_5	The Citizen Digital Assistant eliminates these shortcomings and enables a citizen-controlled communication providing the secure management of digital documents, identities, and credentials.	1564531496	G01.1	Digital assistant
...	...	...	...	...	...
116758	T20.2_2005401280_2	While previous work has concentrated on language support and other platform support, little attention has been placed on graphical user interface variability.	2005401280	T20.2	graphical user interface

3. Results and Evaluation

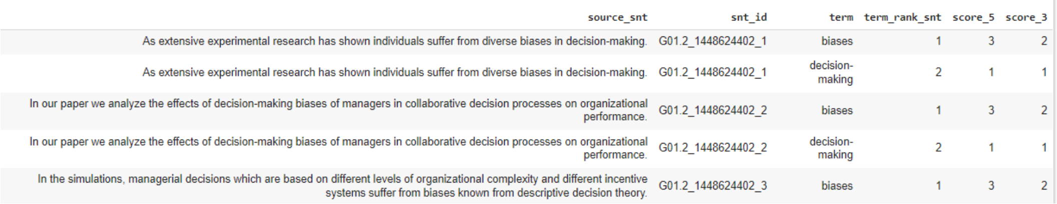
To evaluate our statistical model, we applied it on the training data, since we already have the actual hard terms. We found that the results obtained from the predictive model are similar to the real results. In fact, the following table shows that there are sentences that can have two difficult terms, which is similar to what we obtained.

the result that we found from our predictive model



0	liste_text	difficult term	score_3	score_5
As extensive experimental research has shown individuals suffer from diverse biases in decision-making.	[As, extensive, experimental, research, has, shown, individuals, suffer, from, diverse, biases, in, decision, making]	[biases, decision making]	[2, 1]	[3, 1]
In our paper we analyze the effects of decision-making biases of managers in collaborative decision processes on organizational performance.	[In, our, paper, we, analyze, the, effects, of, decision, making, biases, of, managers, in, collaborative, decision, processes, on, organizational, performance]	[in paper, decision making, biases]	[1, 1, 2]	[1, 1, 3]
In the simulations, managerial decisions which are based on different levels of organizational complexity and different incentive systems suffer from biases known from descriptive decision theory.	[In, the, simulations, managerial, decisions, which, are, based, on, different, levels, of, organizational, complexity, and, different, incentive, systems, suffer, from, biases, known, from, descriptive, decision, theory]	[managerial, incentive, biases]	[1, 1, 2]	[2, 2, 3]
The results illustrate how biases in combination with each other and in different organizational contexts affect organizational performance.	[The, results, illustrate, how, biases, in, combination, with, each, other, and, in, different, organizational, contexts, affect, organizational, performance]	[the results, biases]	[5, 2]	[5, 3]

the true result



source_snt	snt_id	term	term_rank_snt	score_5	score_3
As extensive experimental research has shown individuals suffer from diverse biases in decision-making.	G01.2_1448624402_1	biases	1	3	2
As extensive experimental research has shown individuals suffer from diverse biases in decision-making.	G01.2_1448624402_1	decision-making	2	1	1
In our paper we analyze the effects of decision-making biases of managers in collaborative decision processes on organizational performance.	G01.2_1448624402_2	biases	1	3	2
In our paper we analyze the effects of decision-making biases of managers in collaborative decision processes on organizational performance.	G01.2_1448624402_2	decision-making	2	1	1
In the simulations, managerial decisions which are based on different levels of organizational complexity and different incentive systems suffer from biases known from descriptive decision theory.	G01.2_1448624402_3	biases	1	3	2

4. Conclusion

After evaluating our statistical model (IDF), we found that it performs very well for difficult term extraction. However, the field of difficult term extraction from complex content is currently being improved by trying the latest deep learning methodologies that have been successfully used in other natural language processing tasks and by updating more traditional methodologies. In addition, we are interested in making comparisons between the scores obtained by IDF and Jurassic-1 and making comparisons with evaluation metrics.

References

[1] L. Ermakova, P. Bellot, J. Kamps, D. Nurbakova, I. Ovchinnikova, E. SanJuan, E. Mathurin, R. Hannachi, S. Huet, S. Araujo, Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts, Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022) 13390 (2022).

[2] S. P. Banbury, W. J. Macken, S. Tremblay, D. M. Jones, Auditory distraction and short-term memory: Phenomena and practical implications, Human factors 43 (2001) 12–29.

[3] S. O. L. B. . S. Y. n. Lieber, O., Jurassic-1: Technical details and evaluation, 9.

[4] G. Kavita, What is inverse document frequency (idf)?, 2022.

[5] V. Alto, Understanding pointwise mutual information in nlp, 2020.

[6] B. P. K. J. N. D. O. I. S. E. M. E. A. S. H. R. H. S. . P. N. Ermakova, L., Automatic simplification of scientific texts: Simpletext lab at clef-2022. in m. hagen, s. verberne, c. macdonald, c. seifert, k. balog, k. nørvåg, v. setty (eds.), advances in information retrieval (vol. 13186, pp. 364–373). springer international publishing. https://doi.org/10.1007/978-3-030-99739-7$_4$6(2022).