GR-196 Automatically Improve Writing for Research Papers Using Transformer Based Deep Learning

Abstract

Research papers are generally required to be written in English. It is quite a challenge for non-English speaking researchers to write high quality research papers. Our research focuses on designing, implementing and experimenting with novel transformer based deep learning strategies that can automatically improve the quality of writing for research papers in different domains.

Background

English has become a global language [1]. It is used as a preferred language to communicate in top conferences/journals etc. to share or showcase research. Researchers from across continents write research in their own language and when these works are translated to English, they could be misinterpreted.

Some of the challenges faced by AI translation are [2]:

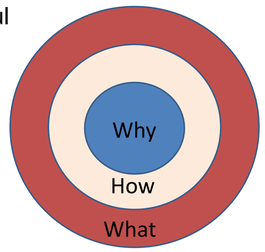
- Misspellings or grammatical mistakes
- Words or phase with multiple meanings or intentions
- Linguistic characteristics like use of tone, metaphors, satire, irony etc. which cannot be define by rules.

Research are being done in this area. We are targeting to find a solution for these problems.

Research Question(s)

Our research focuses on 'The Golden Circles'

1. **Why**: The Purpose
   To provide a helping hand for non-English and English speaking researchers to improvise their academic papers writing style to containing meaningful sentences with relevant words related to their research domain in English that meets both the academics and industry standards.



2. **How**: Strategies and process
   To develop a solution with novel transformer based deep learning strategies that can perform self-supervised learning on large-scale literature repositories and then automatically transform poorly written research papers to professional writings.

3. **What**: The Result
   To develop a product that will be available for all students to improvise their writing style to produce a best quality papers.

Materials and Methods

Designing and implementing novel deep learning strategies based on the SOTA transformer techniques that can perform self-supervised learning on large-scale literature repositories and then automatically transform poorly written research papers to professional writings. Exploratory research is being carried out on Google BERT, T5, and OpenAI GPT-2.

The significant contributions of this project lies in
1. The propose of novel deep learning strategies for this challenging task;
2. Strategies to automatically generating large-scale training data for this challenging task.

Preliminary Results

<Poorly written>

On-line handwriting recognition unusual among sequence labelling tasks in that the underlying generator of the observed data, i.e. the movement of the pen, is recorded directly. However, the raw data can be difficult interpret because each letter is spread over many pen locations. As a consequence, sophisticated pre-processing is required to obtain inputs suitable for conventional sequence labelling algorithms, such as HMMs. In this paper we describe a system capable directly transcribing raw on-line handwriting data. The system consists of a recurrent neural network trained for sequence labelling, combined with a probabilistic language model. In experiments on an unconstrained on-line database, we record excellent results using either raw or pre-processed data, well outperforming a benchmark HMM in both cases.

<Well written>

On-line handwriting recognition is unusual among labeling tasks sequentially in which the underlying generating the observed data, i.e., the movement of the pen, is recorded directly. However, the raw data can be difficult to interpret because each letter is distributed over many locations of the pen. As a result, sophisticated pre-processing is required to obtain adequate for conventional labeling algorithms such as HMMs sequence entries. In this paper a system capable of directly transcribing raw on-line handwriting data. The system consists of a recurrent neural network trained for sequence labelling, combined with a probabilistic language model. In experiments on an unconstrained on-line database, we record excellent results using either raw or pre-processed data, exceeding an HMM benchmark in both cases.

"False Friend" Identification:

A "false friend" is a word which is often confused with a word existing in another language because the two words look or sound similar, but which have different meanings[3]. For instance, the terms "medical device" used in the English version is translated into Spanish and Catalan as producto sanitario ("healthcare product"). The word "medical/medicine" are used in the term "laboratory medicine". We have identified few of these 'false friend' words during translation.

Conclusion

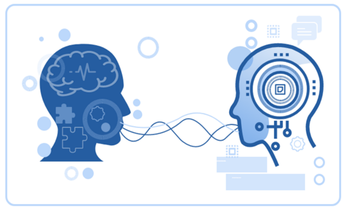
Aim is to find novel strategies to provide a helping hand for non-English speaking researchers to improvise their academic papers writing style containing meaningful sentences with relevant words related to their research domain in English that meets both the academics and industry standards.

Intellectual Merit

The proposed novel deep learning strategies for professional writing using transformers for self supervised learning model and also to generate strategies for large-scale poorly written research papers is beneficial for students to write technical papers/ project reports with meaningful sentences which are acceptable by top conferences/journals provide students to showcase their quality research in a professional manner.

Benefit to Society

Writing papers/project reports etc. are very important tasks to showcase the work done by students and professors. We are providing a helping hand for students to write quality papers in English. This cater to all students irrespective of the countries or languages to write a quality technical writing paper in English using meaningful words used in their area of research to meet the standards of top conferences/journals.



Acknowledgment

I am grateful to Dr. Ying Xie for his support, guidance and valuable expertise in this research domain. I am thankful for his perseverance and motivation to help students like me to excel in school.

Contact Information

Srivarna Settisara Janney
Ph.D. Candidate in Analytics and Data Science
ssettisa@students.kennesaw.edu
Research Advisor: Dr. Ying Xie
Area of Interest: Text Analytics

References

[1] M. Firstova, 'Why has English become a global language? What are the linguistic, political and economical reasons?, Mar 22, 2019.
[2] M.Pisarova, 'what are the biggest challenges facing AI translation technology? July 22, 2022.
[3] Cambridge Dictionaries Online. http://dictionary.cambridge.org/. Accessed 27 January 2013.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv e-prints, October 2018.
[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. Language Models are Unsupervised Multitask Learners, 2019.
[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv:1910.10683, October 2019.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author(s): S.S. Janney
Advisors(s): Dr. Ying Xie