Analyzing a Decade of Evolution: Trends in Natural Language Processing

Anonymous ACL submission

Abstract

Natural Language Processing (NLP) stands at the forefront of the rapidly evolving landscape of Machine Learning, witnessing the emergence and evolution of diverse methodologies over the past decade. This study delves into the dynamic trends within the NLP domain, specifically spanning the years 2010 to 2022, through an empirical analysis of papers presented at conferences hosted by the Association for Computational Linguistics (ACL). Our investigation encompasses an exploration of several key aspects, namely **computational trends**, **research trends** and **geographic trends**. We further investigate the entry cost into NLP, the longevity of hardware and the environmental impact of NLP. [1]

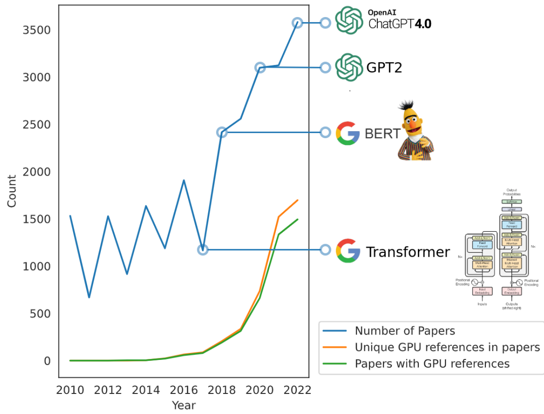

Figure 1: Number of papers published per year in the various ACL conferences, as well as the number of papers with GPU references

1 Introduction

The field of artificial intelligence (AI) has made remarkable strides in recent years, achieving significant milestones across various domains. These advancements range from breakthroughs in computer vision, enabling machines to interpret visual information, to innovations in drug discovery, where AI-driven systems are demonstrating their ability to predict molecular interactions and potential therapeutic effects.

One of the most notable areas of progress is in Natural Language Processing (NLP), where machines have evolved from basic rule-based and statistical systems to sophisticated models capable of understanding, interpreting, and interacting with human language. This evolution, driven by deep learning, seen in **Figure 1**, has revolutionized NLP's capabilities over the past decade.

With the recent introduction of ChatGPT and other open-source Large Language Models (LLMs), the public has begun to recognize the transformative power of these technologies, which is rapidly reshaping how people worldwide work (Eloundou et al., 2023). Although the effects of LLMs have only recently become apparent to the public, many have already experienced their impact. For instance, Google has employed BERT in its search engine since 2019 (Pandu, 2019), and also use NLP techniques to enhance search result comprehension with the use of passages in 2020 (Prabhakar, 2020).

Although the impact of LLMs are mainly considered positive, there are negative aspects, namely the environmental impact to train these models. In 2019, it was estimated that up to 284,019kg of CO_2e (carbon dioxide equivalents) was used to train a transformer using NAS (Strubell et al., 2019). This was later revealed to be overestimated by up to $\times 88$ (Patterson et al., 2021). In 2019, it was estimated by both Nvidia (Leopold, 2019) and Amazon (Barr, 2019) that up to 90% of the workload on machine learning is from inference alone. This iterates that the cost of training a model is, relatively speaking, not that harmful. However, in more recent years, it is becoming increasingly costly to train models. In order to train the Llama

[1]All relevant code and data will be made available on Github upon acceptance.

2 model (Touvron et al., 2023), it was estimated that 539,000kg of CO_2e were used in the training of the model. It was estimated that the 70B model was trained for 1,720,320 GPU hours, using A100 80GB GPUs, equating to 688,128kWh in power consumption assuming a TDP of 400W.

While many papers discuss trends of NLP, they mainly discuss the state of the art (Young et al., 2018; Khurana et al., 2023), tracking how methods and results evolved over time. Additionally, there are various blogs related to upcoming and trending techniques within the field (Insights, 2022; Wolff, 2020). To the best of our knowledge, there is no existing work that attempts to study the various trends using empirical methods. This study aims to address this gap, by examining the key developments in NLP over the last decade. Our analysis is guided by three research questions that aim to uncover the evolving trends in NLP. These questions are:

1. **Computational trends:** What are the dominant computational resources (hardware) in NLP research?
2. **Research trends:** How have NLP tasks, software frameworks and models evolved over the past decade?
3. **Geographic Trends:** How diverse are the publications in the field of NLP?

2 Methodology

To effectively study trends over the past few years, we analyzed papers from top conferences related to Natural Language Processing (NLP). The Association for Computational Linguistics (ACL) stands out as one of the premier conferences for NLP-related research, achieving the highest h5-index in computational linguistics [2], with many of its other events ranking within the top 10 highest h5-indices for computational linguistics. We use the papers from these conferences spanning between the years 2010 and 2022 to perform analysis (detailed in **Appendix A.1**). The remaining methods is briefly summarized in **Figure 2**.

We download the papers in a PDF format, which is considered an unstructured format. When extracting text from a PDF, this can lead to the inclusion of unwanted data. The simpler approach

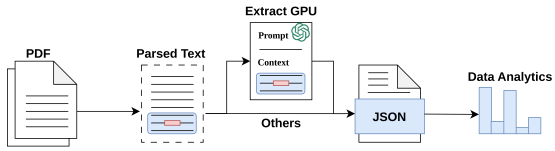

Figure 2: Overview of the methodology used to extract information

involves extracting all text from a PDF in an unstructured format, using PyPDF2 [3]. Alternatively, a more structured approach attempts to semantically parse the PDF data into sections visible to a viewer, using the SciPDF Parser [4]. However, the structured approach comes with the disadvantage of potential data loss, whereas the unstructured approach sacrifices structure and may include data that can interfere with subsequent data analysis. In this work, we employ both approaches. Further details regarding the implementation of these two approaches can be seen in the **Appendix A.2**.

We further collect citation information of all papers, with the use of Semantic Scholar API (Kinney et al., 2023) (this was performed on July 24, 2022). When performing the match for citation results, we use Jaro similarity and set a threshold of 7.5 to account for inconsistencies of paper names. We were unable to find citation information for 20 of the papers.

The next step involved searching for specific keywords in the text that we aimed to extract. To successfully mine information about GPUs, we employed a two-stage pipeline. The initial step focused on identifying general GPU architectures in the text using keywords such as 'rtx', 'gpu', 'nvidia', 'tesla', 'quadro', 'geforce' and 'gtx', which is detailed further in **Appendix A.3**. Utilizing these keywords, we extracted a context of up to 500 characters surrounding the supporting word. This was achieved by chunking data into sentences and recursively adding these sentences to the context until the character limit was reached. We then aimed at extracting the exact GPU used, using exact dictionary matching.

Subsequently, we extracted exact GPU information from the context using ChatGPT (detailed in **Appendix A.4**). Pre-processing was performed on the GPU names, removing all keywords mentioned earlier. Upon analyzing the data, we opted

[2] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics

[3] https://pypi.org/project/PyPDF2/

[4] https://github.com/titipata/scipdf_parser

for ChatGPT's annotations over exact matching.

We also collected some statistics regarding frameworks used, models used and general tasks authors aimed at solving. These were done on a simple dictionary matching scheme, whereby in cases where it makes sense, spaces were added/removed to maximize correct matches. We also limited the sections to analyze in the cases of architectures and NLP tasks, whereas for the frameworks, the entire text is searched.

3 Results

A summary of some general statistics are seen in **Figure 1**, where the overall number of papers collected is presented, along with the number of papers with GPU references and the number of unique GPUs referenced in each paper. Analyzing first the number of papers, we notice some oscillation in the beginning, related to biannual conferences, followed by an explosion in NLP papers from around 2018. In 2022 almost half the papers have mentioned specific GPUs used, which indicates that access to GPUs may be a limiting factor for publication in this field. We further note that only in more recent years do we see papers using multiple GPU architectures. In total 25,591 papers were collected of which 5,961 contain GPUs. In more recent years, the NVIDIA V100 emerged as the most popular GPU (see **Figure 5** in the appendix), with newer GPUs such as the A100 and 3090 appearing to be on the rise. While older GPUs have a diminishing presence in later years, whereas the counts of newer GPUs surge, indicating the transition to updated hardware. This shift is likely attributed to the escalating memory requirements essential for performing NLP.

Analyzing the most popular frameworks used (see **Figure 6** in the appendix), it is unsurprising that PyTorch and Hugging Face lead the field, primarily owing to their accessible APIs, facilitating rapid development. In close pursuit is TensorFlow, which has gradually lost popularity with NLP researchers, in more recent years. Overall, the framework counts are lower than the GPU frequencies, indicating that the frameworks used are not highly discussed. The substantial surge in Hugging Face's prevalence may also be attributed to the utilization of footnotes, as numerous authors reference models from the Hugging Face Hub in their text.

We further analyzed the most popular algorithms and techniques presented in the conferences (see **Figure 3**). In the earlier years, there was a significant focus on Bayesian-based techniques with Support Vector Machines (SVMs). However, around 2013, the field shifted towards deep learning, and the use of neural networks and LSTMs emerged as dominant algorithms for solving tasks. We further date this to the use of embeddings such as Glove (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013). Subsequently, there was an explosion in 2018 with the introduction of the transformer architecture, particularly BERT (Devlin et al., 2019), and later GPT in 2020 (Radford et al., 2018, 2019; Brown et al., 2020), which we expect to further grow within upcoming years.

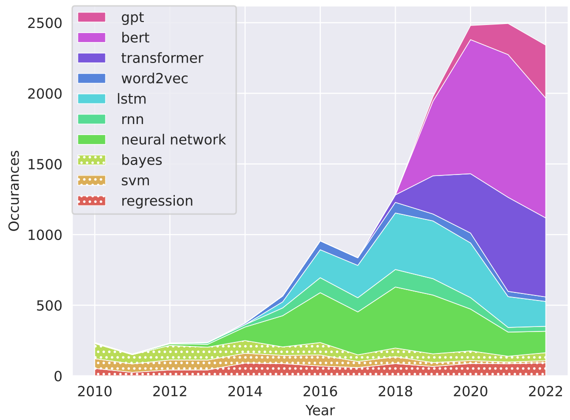

Figure 3: Most popular NLP models, (Classical machine learning approaches are marked with dots, deep learning approaches are unmarked)

Examining the most popular tasks (**Figure 4**), it is evident that translation is the predominant task under investigation. Up until 2018, most trends remained consistent. However, with the introduction of transformer-based architectures, there is a substantial surge in the exploration of question answering tasks as well as text generation tasks, leading to significant progress in these tasks.

Following this, we analyzed the data on a per-country basis. To determine the country of each paper, extracted from the paper using the structured reader. Out of 25,559 papers, we were able to extract country information for 13,365 papers. With this information, we plotted heatmaps showcasing the countries with the most papers published (**Figure 10**), citations (**Figure 11**), and the number of citations per country (**Figure 12**), all **Figures** are in the appendix. The countries with the most papers are China (3,141), the United States (2,538), and Germany (1,880). However, when comparing this value to the average number of citations per

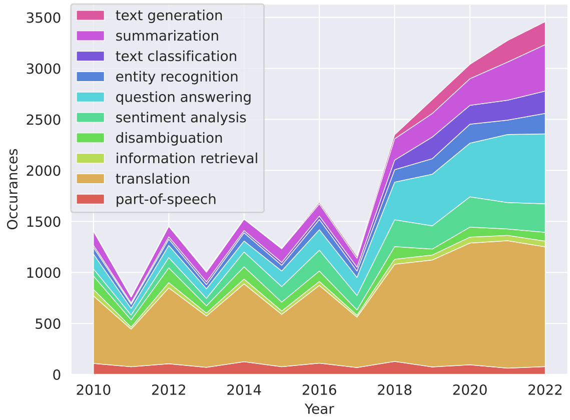

Figure 4: Most popular NLP tasks

paper per country, the results differ, with the top countries being Mexico (101.66), Colombia (69.8), and the United States (67.72). If our data is correct, this demonstrates that relevant research remains impactful irrespective of the country of origin.

We further estimated the environmental impact of NLP. We estimate that the power usage needed by researchers to produce publications for the year 2022, is around 733,104kWh. This assumes that for each paper with a gpu reference, the GPU is running for 1 month at 50% power usage, and that average power draw of the server is 1200W. Further details regarding these assumptions can be seen in the appendix. This value slightly exceeds the power needed to train Llama 2 70B. Although this is a naive estimate, it shows that the energy usage of all research, barely compares to that used in large companies where foundational models are being trained.

4 Discussion

In this section, we aim to address the various questions posed at the beginning of the paper. Regarding **Computational Trends**, we observe an increasing entry cost into the field of NLP, as many papers now require expensive dedicated hardware. However, there still appears to be scope for research without such costly hardware requirements. Investing significantly in hardware raises questions about its longevity, as some older hardware has become mostly irrelevant. Nevertheless, newer hardware with increased VRAM should remain pertinent unless models rapidly grow in size. For example the Nvidia V100 still remains relevant given its age, due to its large VRAM. One limiting factor in research is the availability of GPU manufacturers releasing GPUs with significantly increased VRAM. Addressing the environmental impact of training models, as discussed by previous authors, the relative environmental impact of the field does not seem large. The impact of an average researcher pales in comparison to that of larger companies.

Discussing **Research Trends**, the introduction of the transformer architecture has significantly impacted the field, fostering positive growth and enabling research on more challenging topics. We anticipate that the use of ChatGPT will further contribute to the rising number of papers. Researchers are likely to conduct studies without dedicated hardware, utilizing ChatGPT's API and various other APIs released within the last year. Currently, Hugging Face stands out as the most used software in NLP, followed by PyTorch.

Finally, in investigating **Geographic Trends**, regardless of the country of origin, papers with substantive content have the potential to achieve high impact. While certain countries may have a higher number of accepted papers, the determining factor for impact remains the quality of the paper's contents.

5 Conclusion

In this study, we conducted an analysis of Natural Language Processing (NLP) trends from 2010 to 2022, focusing on ACL based conference papers. Our findings highlight an increasing entry cost to NLP, driven by the demand for expensive hardware. Despite uncertainties about hardware longevity, newer high VRAM options suggest potential stability. The environmental impact of NLP training appears relatively modest, with larger companies overshadowing individual researchers. The impact of the transformer architecture on **Research Trends** has driven increased output and exploration of complex topics. Anticipating continued impact, we foresee a rise in papers facilitated by tools like ChatGPT and other APIs, enabling research without dedicated hardware. Hugging Face and PyTorch currently dominate NLP software. In our analysis of **Geographic Trends**, we note that the primary determinant of impact is the paper's quality, and does not appear to be limited by country of origin. To conclude, our study provides insights into the evolving NLP landscape, briefly overviewing trends present in the last decade of research.

6 Limitations

The primary limitation of this work lies in the automatic extraction of GPUs used in papers. We acknowledge that this value may be underestimated since many papers do not include this information within the text. Despite this potential underestimation, we are confident that if a GPU was mentioned within an extracted context, it was almost always correctly identified, enhancing the reliability of this study.

As mentioned earlier, we desired to include information about the training time of algorithms to more accurately calculate the energy consumption of model training. However, we were unable to confidently extract this information for publication. While we naively estimate the power consumption of machine learning models, this value cannot be confidently estimated even with the training time of algorithms. We lack information about the time spent on prototyping beforehand, making any estimate regarding computing time inherently inaccurate.

In **Figures** 3 and 4, the lists of tasks and algorithms may not be exhaustive. However, to the best of our ability, we ensured that all significant tasks and algorithms were included. Various other algorithms were tested but deemed non-significant and subsequently removed, including 'tf-idf', 'nn', 'random forest', 'knn', 'recurrent neural network', 'pca', 'rbf', and 'lda'.

References

Jeff Barr. 2019. Amazon EC2 Update – Inf1 Instances with AWS Inferentia Chips for High Performance Cost-Effective Inferencing | AWS News Blog. Amazon Blog.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models.

Google. 2023. Efficiency – Data Centers – Google.

StartUs Insights. 2022. 9 Natural Language Processing Trends in 2023.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

George Leopold. 2019. AWS to Offer Nvidia's T4 GPUs for AI Inferencing.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nayak Pandu. 2019. Understanding searches better than ever before. Google Blog.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Raghavan Prabhakar. 2020. How AI is powering a more helpful Google. Google Blog.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Petroc Taylor. 2023. Data center average annual PUE worldwide 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Rachel Wolff. 2020. 9 Natural Language Processing (NLP) Trends in 2022.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75.

A Appendix

A.1 Conference Information

From each conference, we exclusively download the main proceedings, excluding workshops, demonstrations, tutorials, student sessions, industry events, etc. This results in a dataset comprising the main proceedings of the conferences, encompassing both long and short papers. The List of conferences used in this study are as follows:

1. Annual Meeting of the Association for Computational Linguistics (ACL)
2. Conference on Empirical Methods in Natural Language Processing (EMNLP)
3. North American Chapter of the Association for Computational Linguistics (NAACL)
4. International Conference on Computational Linguistics (COLING)
5. International Conference on Language Resources and Evaluation (LREC)
6. Conference on Computational Natural Language Learning (CoNLL)
7. European Chapter of the Association for Computational Linguistics (EACL)
8. International Joint Conference on Natural Language Processing (IJCNLP)

Furhter information reagarding when the various conferences were held can be seen in **Table 1**.

A.2 PDF Parsing

We provide more information regarding the PDF parsing and processing:

- Unstructured Reader: For the unstructured reader, we utilized the popular Python PDF reader, PyPDF2 [5]. This library enables us to extract all text on a page. In a PDF, there is no inherent definition of a line break, so at the end of each line, a line break is manually inserted. To address this, we use regular expressions to remove various line breaks and replace them with spaces. Similarly, in PDFs, when a word extends past the natural width of the page, the word is broken with a hyphen. This is also corrected with regular expressions.

  Next, page numbers are removed from the text, which can appear as either the first or last characters on the page string. Finally, a sentence tokenizer is applied to the text to split it into sentences, to allow for some structure. This process results in relatively clean text, although certain aspects of the text may remain uncleaned.

- Structured Reader: For structured data, we employed the SciPDF Parser [6], built on the library, Generation of Bibliographic Data (GROBID) [7], which utilizes machine learning for restructuring PDFs. Although this approach is often imperfect and may contain errors in its structuring, it allows for some degree of structuring within the documents.

A.3 GPU selection and pre-processing

In order to perform exact dictionary matching on the GPUs, we built a dictionary from two different sources [8] [9], using additional annotation by hand to ensure the entries are correct. We needed to perform various pre-processing steps to contain a dictionary entry that is compact, due to the amount of different ways to represent a GPU, for example 'Nvidia GeForce RTX 3080' and 'Nvidia 3080 GPU' would both reference the same GPU, which would be matched with only '3080'. This involves adding spaces and removing hyphens where applicable.

A.4 ChatGPT information

With regards to our prompt, our initial idea was to extract various different sources of information into a single JSON field. In order to do this we used the following prompt:

> "You are a machine learning expert. Your goal is to extract correct information from a given CONTEXT and answer the QUESTION correctly. When in doubt, use the value -1.
> **CONTEXT:** CONTEXT
> **QUESTION:** What is the total training time of the models, explaining reasoning, return only a JSON: {total_time: NUMBER, unit: MINUTE/HOURS/DAYS,

[5] https://pypi.org/project/PyPDF2/

[6] https://github.com/titipata/scipdf_parser

[7] https://github.com/kermitt2/grobid

[8] https://developer.nvidia.com/cuda-gpus

[9] https://www.techpowerup.com/gpu-specs

Conference	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
ACL	x	x	x	x	x	x	x	x	x	x	x	x	x
COLING	x		x		x		x		x		x		x
CoNLL	x	x	x	x	x	x	x	x	x	x	x	x	x
EACL			x		x			x				x	x
EMNLP	x	x	x	x	x	x	x	x	x	x	x	x	x
IJCNLP		x		x		x				x		x	x
LREC	x		x		x		x		x		x		x
NAACL	x		x	x		x	x		x	x		x	x

Table 1: Conference years for ACL-related events.

gpus:[{gpu: GPU_NAME, number_of_gpus: NUMBER},...]}"

We attempted to extract information regarding the training time of the models, however this information was only correct in some cases, which is why we did not further investigate this. Similarly, we attempted to extract the number of GPUs used for training, which was correctly estimated when the number of GPUs is explicitly mentioned. However, we struggled to keep this information when merging data from the same source, and this could lead to incorrectly estimating the number of GPUs used. For example, in one context for a paper, it could mention that 4 GPUs were used for training, and later in the paper, it might mention only 1 GPU. This is hard to account for, as it can be seen as either 4 GPUs or 5 GPUs.

We then compared the data extracted from ChatGPT with the data we manually extracted from the dictionary extraction methods. In approximately 80% of the cases, the matching was equal between the two methods. In the remaining cases, we empirically identified ChatGPT as superior. This was mainly due to instances where hyperparameter values were often mistaken for older GPU names such as ‘2000’ and ‘680M’. This is why we used the data extracted form the ChatGPT analysis in the paper.

A.5 Supplementary Results

Taking a closer look at the number of GPUs identified in papers (**Figure 5**), where the GPUs selected were among the top 3 for each year. Examining the overall trend in the data, a clear increase in the number of GPUs detected in papers per year is evident, signifying exponential growth. Although this may not constitute a direct comparison, as earlier papers might not specify the architecture, this data reveals a rising demand for GPUs.

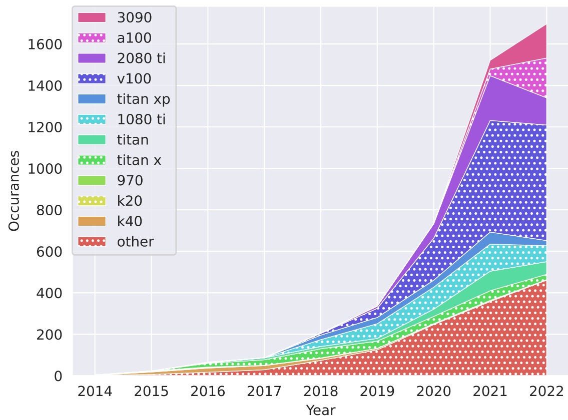

Figure 5: Count of the top 3 GPUs from each year

Figure 6 showcases the most popular NLP frameworks. It is clear that Hugging Face has become the dominant framework for NLP based research in the more recent years.

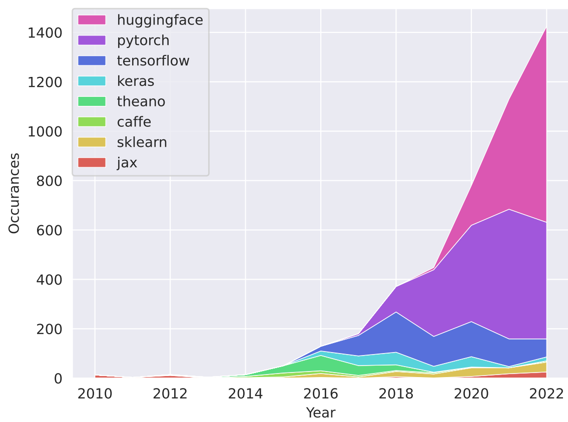

Figure 6: Most popular NLP frameworks

In Figures 7 and 8, we observe the number of papers and citations from each conference per year. Beginning with the number of papers produced at each conference, there is a noticeable increase in the quantity of papers published in most conferences, while LREC remains relatively constant. The high value for NAACL in 2019, is due to the publication of the paper: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Subsequently, there is a decline in the number of citations after 2019, which can be attributed to newer works having fewer citations due to having less 'visible' time.

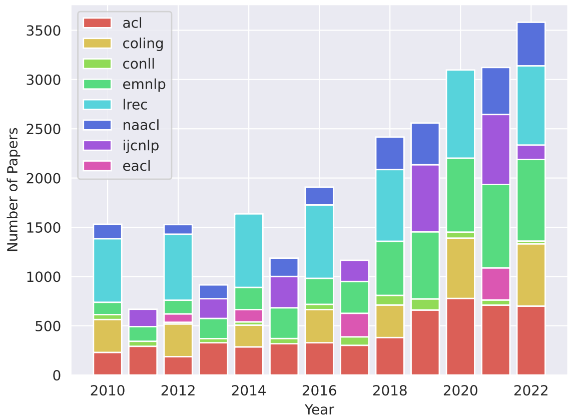

Figure 7: Number of papers published in each conference per year

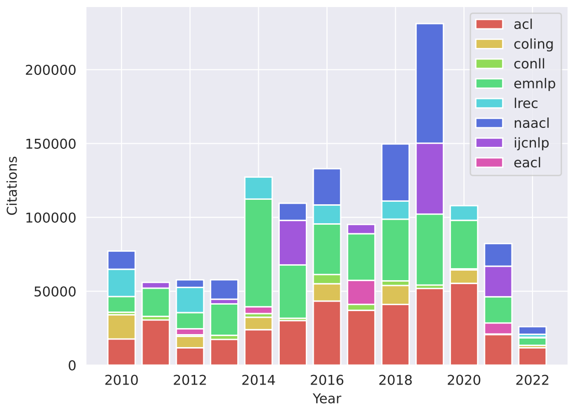

Figure 8: Number of citations from each paper published in each conference per year

Examining the table containing the Top 10 most cited papers (**Table 2**), we observe that 6 out of 10 of the top papers include contributions from EMNLP. Notably, the third and fourth most popular papers are from 2014 as well. Comparing these values with **Figure 8**, the results remain consistent, with ACL having a higher average number of citations in most cases(**Figure 9**), corresponding to the various metrics used to rank these conferences.

Following this, we analyzed the data on a per-

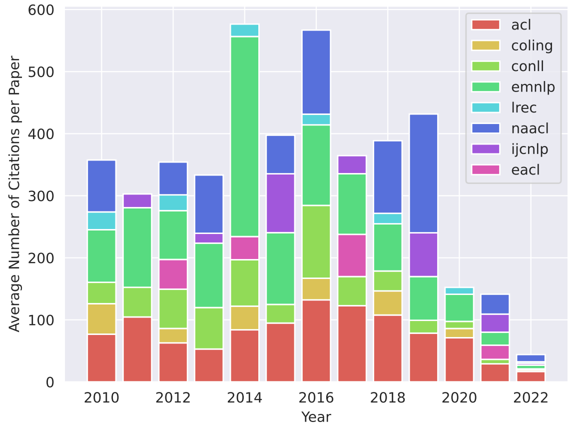

Figure 9: Average number of citations per paper, per conference, per year

country basis. To determine the country of each paper, extracted from the paper using the structured reader. With this information, we plotted heatmaps showcasing the countries with the most papers published (**Figure 10**), citations (**Figure 11**), and the number of citations per country (**Figure 12**). The countries with the most papers are China (3,141), the United States (2,538), Germany (1,880), the United Kingdom (1,237), France (909), and Japan (855). However, when comparing this value to the average number of citations per paper per country, the results differ, with the top countries being Mexico (101.66), Colombia (69.8), the United States (67.72), Israel (60), and Germany (57.63).

Turning to the environmental impact of NLP let's consider an example. In 2022, there were 557 mentions of the v100 GPU, which has a Thermal Design Power (TDP) of 300W. TDP represents the maximum heat the GPU can generate under sustained workload conditions and is utilized here as an estimate of power draw. Assuming a model is trained, on average, for 1 week with GPUs running at 70% usage, the power consumption, as per **Table 3**, would be $35.28 \times 557 = 19650.96$kWh. This estimate is simplistic, covering only the GPU's power consumption, excluding the server's and data center's power usage. The latter is commonly estimated by the Power Usage Effectiveness (PUE) coefficient, estimated at 1.58 (Taylor, 2023) (Google reported a PUE of 1.10 in 2023 (Google, 2023)). Additionally, this does not account for models trained on multiple GPUs. For a more realistic estimate, assuming a TDP of 1200W, the consumption would be $141.7 \times 557 = 78,926.9$kWh.

Extending this to the total GPUs, assuming each

Paper Title	Year	Conference	Citations
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	2019	NAACL	56,293
GloVe: Global Vectors for Word Representation	2014	EMNLP	27,236
Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation	2014	EMNLP	18,698
Convolutional Neural Networks for Sentence Classification	2014	EMNLP	11,880
Deep Contextualized Word Representations	2018	NAACL	9,800
Effective Approaches to Attention-based Neural Machine Translation	2015	EMNLP	6,939
Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank	2013	EMNLP	6,601
Neural Machine Translation of Rare Words with Subword Units	2016	ACL	6,024
SQuAD: 100,000+ Questions for Machine Comprehension of Text	2016	EMNLP	5,793
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension	2020	ACL	5,391

Table 2: Top 10 cited papers from ACL conferences.

GPU averages a TDP of 300W with the server averaging 1200W, with 1697 GPU references in papers from 2022, and prototyping/training done for 1 month at 50% power usage, this would imply $1697 \times 432 = 733,104$kWh.

TDP	Hours	Average usage (kWh)			
		40%	50%	70%	100%
300w	1	0.12	0.15	0.21	0.30
	1 * 24	2.88	3.60	5.04	7.20
	7 * 24	20.16	25.20	35.28	50.40
	30 * 24	86.40	108.00	151.20	216.00
600w	1	0.24	0.30	0.42	0.60
	1 * 24	5.76	7.20	10.08	14.40
	7 * 24	40.32	50.40	70.56	100.80
	30 * 24	172.80	216.00	302.40	432.00
1200w	1	0.48	0.60	0.84	1.20
	1 * 24	11.52	14.40	20.16	28.80
	7 * 24	80.64	100.80	141.12	201.60
	30 * 24	345.60	432.00	604.80	864.00
2400w	1	0.96	1.20	1.68	2.40
	1 * 24	23.04	28.80	40.32	57.60
	7 * 24	161.28	201.60	282.24	403.20
	30 * 24	691.20	864.00	1209.60	1728.00

Table 3: Average power consumption (kWh)

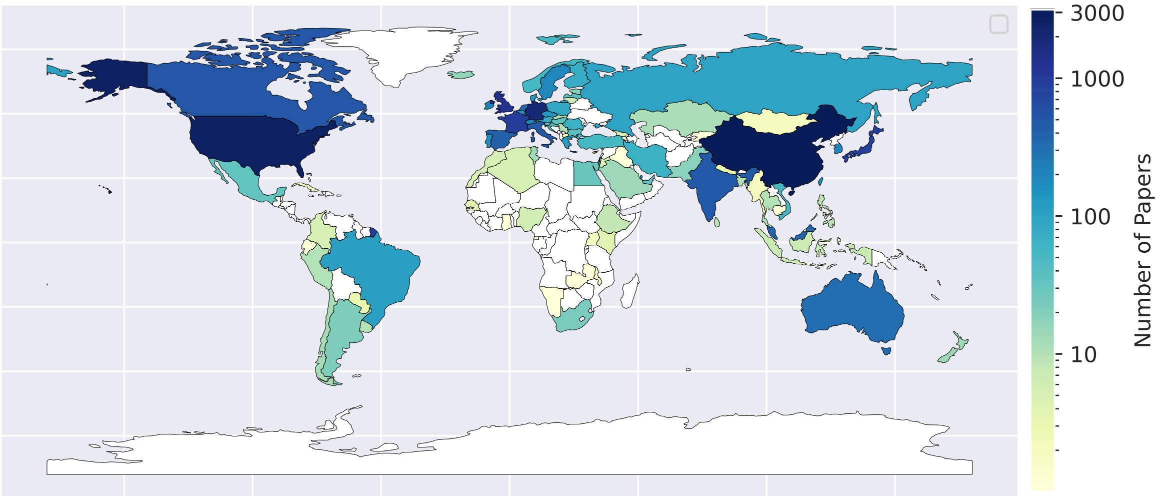

Figure 10: Number of Papers published per country

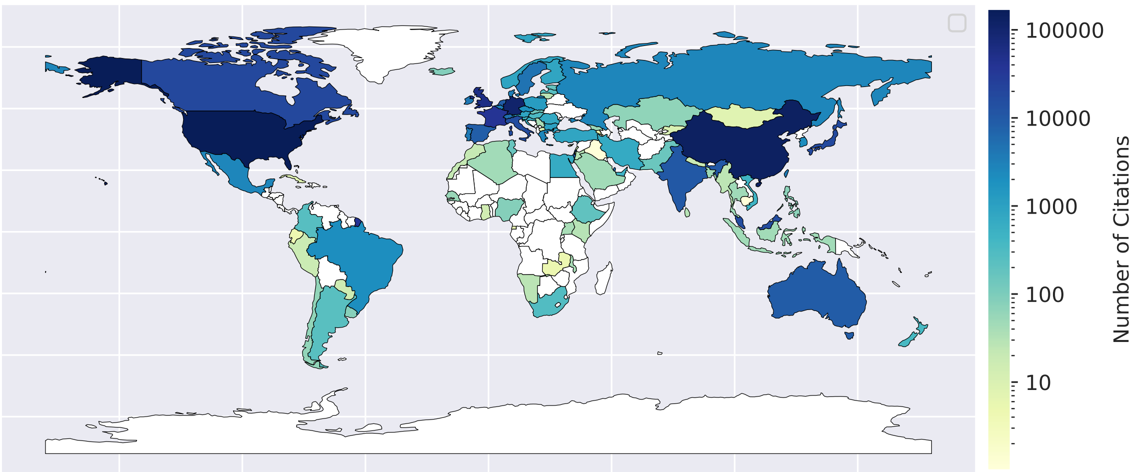

Figure 11: Number of citations per country

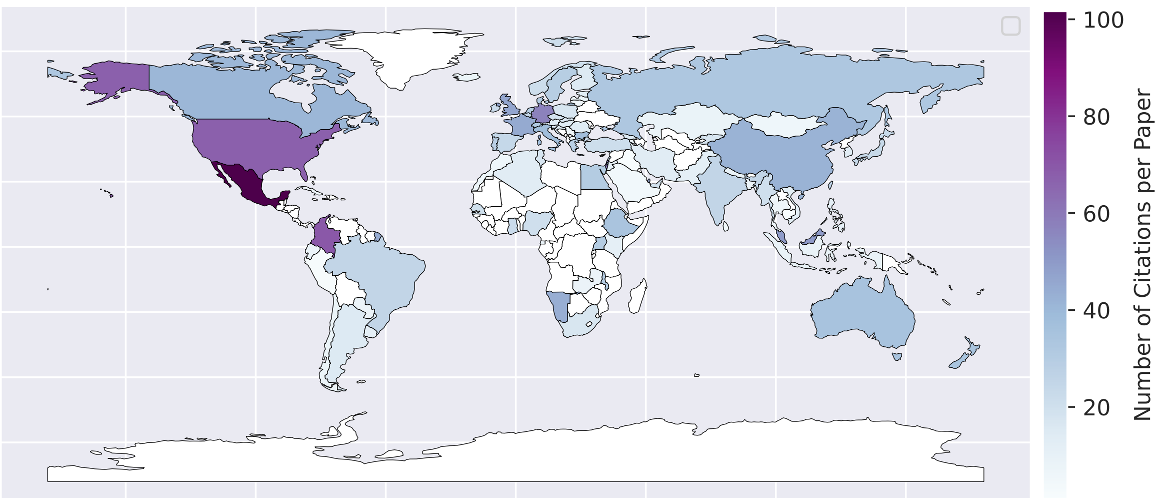

Figure 12: Average number of citations per paper per country