This post is based on our RANLP 2023 paper "Exploring the Landscape of Natural Language Processing Research". You can read more details there.

Introduction

As an efficient approach to understand, generate, and process natural language texts, research in natural language processing (NLP) has exhibited a rapid spread and wide adoption in recent years. Given the rapid developments in NLP, obtaining an overview of the domain and maintaining it is difficult. This blog post aims to provide a structured overview of different fields of study in NLP and analyzes recent trends in this domain.

> *Fields of study are academic disciplines and concepts that usually consist of (but are not limited to) tasks or techniques.*

In this article, we investigate the following questions:

- ***What are the different fields of study investigated in NLP?***
- ***What are the characteristics and developments over time of the research literature in NLP?***
- ***What are the current trends and directions of future work in NLP?***

Although most fields of study in NLP are well-known and defined, there currently exists no commonly used taxonomy or categorization scheme that attempts to collect and structure these fields of study in a consistent and understandable format. Therefore, getting an overview of the entire field of NLP research is difficult. While there are lists of NLP topics in conferences and textbooks, they tend to vary considerably and are often either too broad or too specialized. Therefore, we developed a taxonomy encompassing a wide range of different fields of study in NLP. Although this taxonomy may not include all possible NLP concepts, it covers a wide range of the most popular fields of study, whereby missing fields of study may be considered as subtopics of the included fields of study. While developing the taxonomy, we found that certain lower-level fields of study had to be assigned to multiple higher-level fields of study rather than just one. Therefore, some fields of study are listed multiple times in the NLP taxonomy, but assigned to different higher-level fields of study. The final taxonomy was developed empirically in an iterative process together with domain experts.

The taxonomy serves as an overarching classification scheme in which NLP publications can be classified according to at least one of the included fields of study, even if they do not directly address one of the fields of study, but only subtopics thereof. To analyze recent developments in NLP, we trained a weakly supervised model to classify ACL Anthology papers according to the NLP taxonomy.

You can read more details about the development process of the classification model and the NLP taxonomy in our paper.

Different fields of study in NLP 📖

The following section provides short explanations of the fields of study concepts included in the NLP taxonomy above.

Multimodality

Multimodality refers to the capability of a system or method to process input of different types or *modalities* (Garg et al., 2022). We distinguish between systems that can process text in natural language along with **visual data**, **speech & audio**, **programming languages**, or **structured data** such as tables or graphs.

Natural Language Interfaces

Natural language interfaces can process data based on natural language queries (Voigt et al., 2021), usually implemented as **question answering** or **dialogue & conversational systems**.

Semantic Text Processing

This high-level field of study includes all types of concepts that attempt to derive meaning from natural language and enable machines to interpret textual data semantically. One of the most powerful fields of study in this regard are **language models** that attempt to learn the joint probability function of sequences of words (Bengio et al., 2000). Recent advances in language model training have enabled these models to successfully perform various downstream NLP tasks (Soni et al., 2022). In **representation learning**, semantic text representations are usually learned in the form of embeddings (Fu et al., 2022), which can be used to compare the **semantic similarity** of texts in **semantic search** settings (Reimers and Gurevych, 2019). Additionally, **knowledge representations**, e.g., in the form of knowledge graphs, can be incorporated to improve various NLP tasks (Schneider et al., 2022).

Sentiment Analysis

Sentiment analysis attempts to identify and extract subjective information from texts (Wankhade et al., 2022). Usually, studies focus on extracting **opinions**, **emotions**, or **polarity** from texts. More recently, **aspect-based sentiment analysis** emerged as a way to provide more detailed information than general sentiment analysis, as it aims to predict the sentiment polarities of given aspects or entities in text (Xue and Li, 2018).

Syntactic Text Processing

This high-level field of study aims at analyzing the grammatical syntax and vocabulary of texts (Bessmertny et al., 2016). Representative tasks in this context are **syntactic parsing** of word dependencies in sentences, **tagging** of words to their respective part-of-speech, **segmentation** of texts into coherent sections, or **correction of erroneous texts** with respect to grammar and spelling.

Linguistics & Cognitive NLP

Linguistics & Cognitive NLP deals with natural language based on the assumptions that our linguistic abilities are firmly rooted in our cognitive abilities, that meaning is essentially conceptualization, and that grammar is shaped by usage (Dabrowska and Divjak, 2015). Many different **linguistic theories** are present that generally argue that language acquisition is governed by universal grammatical rules that are common to all typically developing humans (Wise and Sevcik, 2017). **Psycholinguistics** attempts to model how a human brain acquires and

produces language, processes it, comprehends it, and provides feedback (Balamurugan, 2018). **Cognitive modeling** is concerned with modeling and simulating human cognitive processes in various forms, particularly in a computational or mathematical form (Sun, 2020).

Responsible & Trustworthy NLP

Responsible & trustworthy NLP is concerned with implementing methods that focus on fairness, **explainability**, accountability, and **ethical** aspects at its core (Barredo Arrieta et al., 2020). **Green & sustainable NLP** is mainly focused on efficient approaches for text processing, while **low-resource NLP** aims to perform NLP tasks when data is scarce. Additionally, **robustness in NLP** attempts to develop models that are insensitive to biases, resistant to data perturbations, and reliable for out-of-distribution predictions.

Reasoning

Reasoning enables machines to draw logical conclusions and derive new knowledge based on the information available to them, using techniques such as deduction and induction. **Argument mining** automatically identifies and extracts the structure of inference and reasoning expressed as arguments presented in natural language texts (Lawrence and Reed, 2019). **Textual inference**, usually modeled as entailment problem, automatically determines whether a natural-language *hypothesis* can be inferred from a given *premise* (MacCartney and Manning, 2007). **Commonsense reasoning** bridges premises and hypotheses using world knowledge that is not explicitly provided in the text (Ponti et al., 2020), while **numerical reasoning** performs arithmetic operations (Al-Negheimish et al., 2021). **Machine reading comprehension** aims to teach machines to determine the correct answers to questions based on a given passage (Zhang et al., 2021).

Multilinguality

Multilinguality tackles all types of NLP tasks that involve more than one natural language and is conventionally studied in **machine translation**.
Additionally, **code-switching** freely interchanges multiple languages within a single sentence or between sentences (Diwan et al., 2021), while **cross-lingual transfer** techniques use data and models available for one language to solve NLP tasks in another language.

Information Retrieval

Information retrieval is concerned with finding texts that satisfy an information need from within large collections (Manning et al., 2008). Typically, this involves retrieving **documents** or **passages**.

Information Extraction & Text Mining

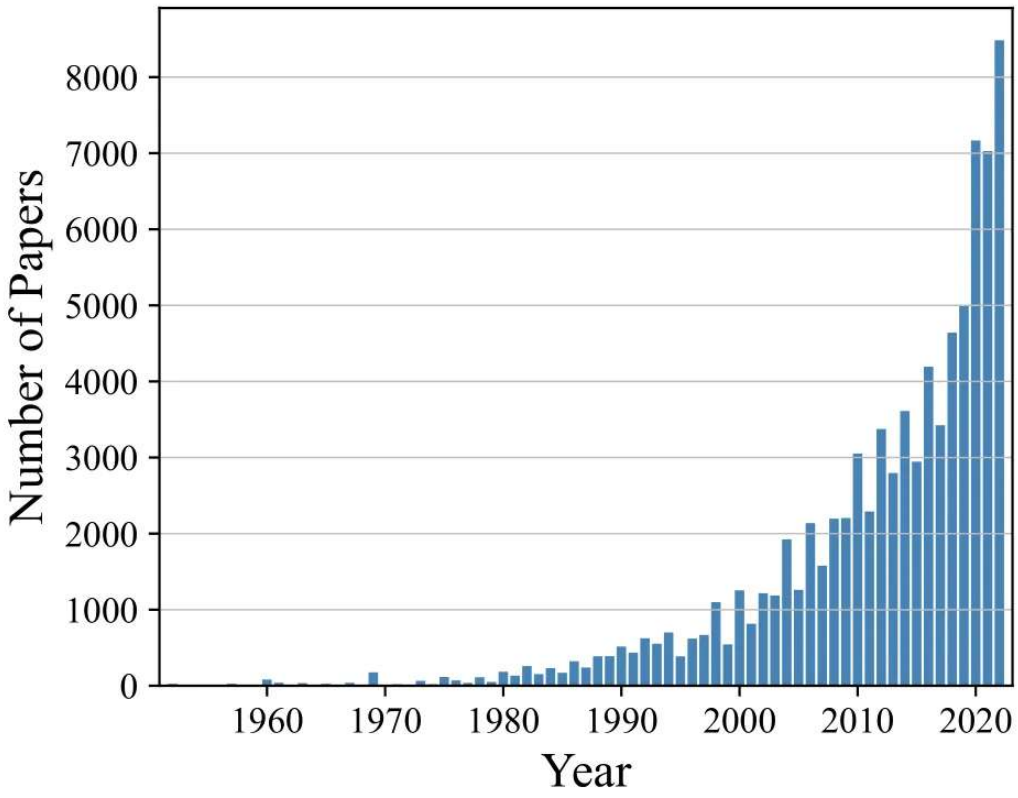
This field of study focuses on extracting structured knowledge from unstructured text and enables the analysis and identification of patterns or correlations in data (Hassani et al., 2020). **Text classification** automatically categorizes texts into predefined classes (Schopf et al., 2021), while **topic modeling** aims to discover latent topics in document collections (Grootendorst, 2022), often using **text clustering** techniques that organize semantically similar texts into the same clusters. **Summarization** produces summaries of texts that include the key points of the input in less space and keep repetition to a minimum (El-Kassas et al., 2021). Additionally, the information extraction & text mining field of study also

includes **named entity recognition**, which deals with the identification and categorization of named entities (Leitner et al., 2020), **coreference resolution,** which aims to identify all references to the same entity in discourse (Yin et al., 2021), **term extraction,** which aims to extract relevant terms such as keywords or keyphrases (Rigouts Terryn et al., 2020), **relation extraction** that aims to extract relations between entities, and **open information extraction** that facilitates the domain-independent discovery of relational tuples (Yates et al., 2007).

Text Generation

The objective of text generation approaches is to generate texts that are both comprehensible to humans and indistinguishable from text authored by humans. Accordingly, the input usually consists of text, such as in **paraphrasing** that renders the text input in a different surface form while preserving the semantics (Niu et al., 2021), **question generation** that aims to generate a fluid and relevant question given a passage and a target answer (Song et al., 2018), or **dialogue-response generation** which aims to generate natural-looking text relevant to the prompt (Zhang et al., 2020). In many cases, however, the text is generated as a result of input from other modalities, such as in the case of **data-to-text generation** that generates text based on structured data such as tables or graphs (Kale and Rastogi, 2020), **captioning** of images or videos, or **speech recognition** that transcribes a speech waveform into text (Baevski et al., 2022).

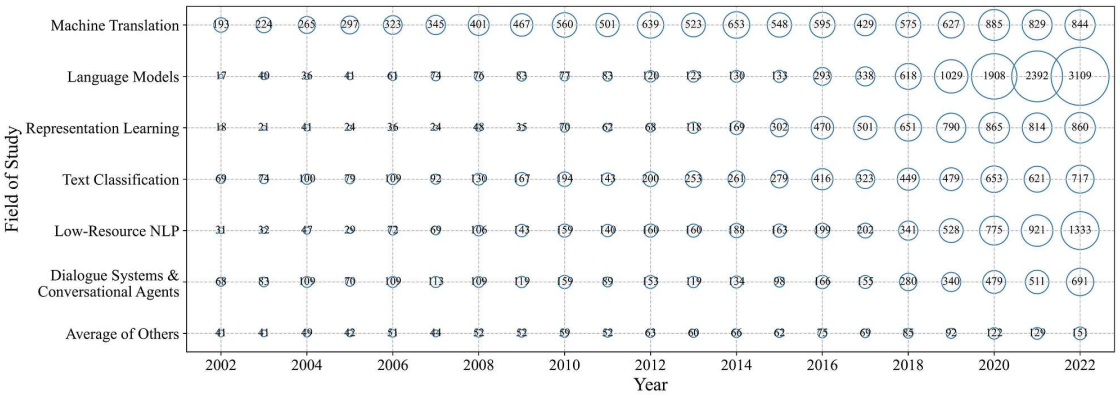
Characteristics and Developments in NLP



The number of papers per year in the ACL Anthology from 1952 to 2022. Image by author

Considering the literature on NLP, we start our analysis with the number of studies as an indicator of research interest. The distribution of publications over the 50-year observation period is shown in the Figure above. While the first publications

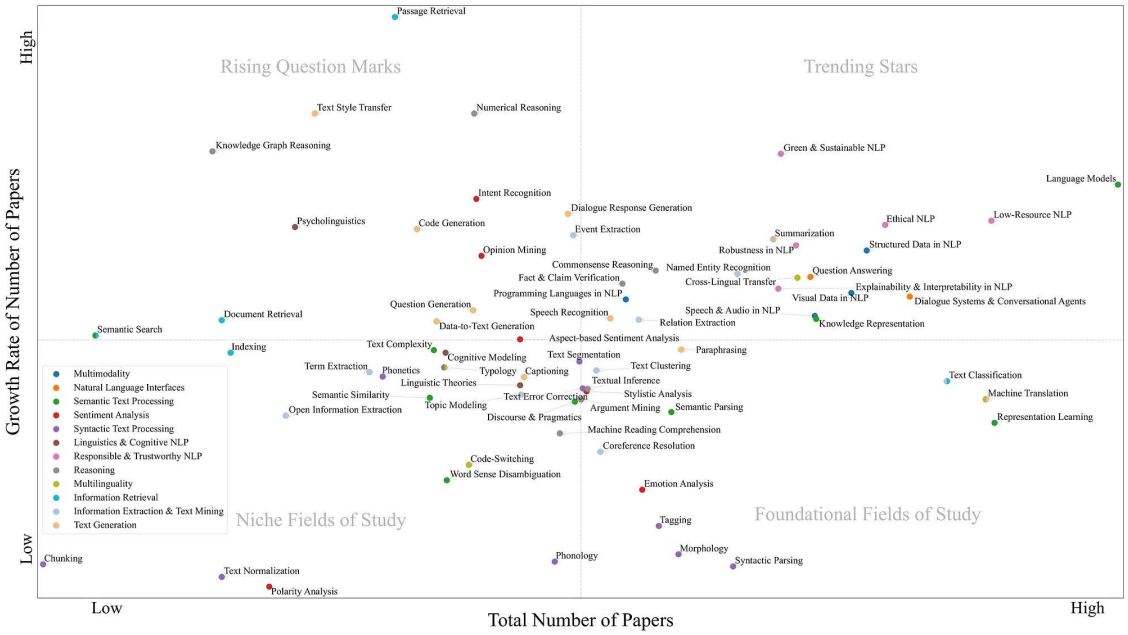
appeared in 1952, the number of annual publications grew slowly until 2000. Accordingly, between 2000 and 2017, the number of publications roughly quadrupled, whereas in the subsequent five years, it has doubled again. We therefore observe a near-exponential growth in the number of NLP studies, indicating increasing attention from the research community.



Distribution of the number of papers by most popular fields of study from 2002 to 2022. Image by author.

Examining the figure above, the most popular fields of study in the NLP literature and their recent development over time are revealed. While the majority of studies in NLP are related to *machine translation* or *language models*, the developments of both fields of study are different. *Machine translation* is a thoroughly researched field that has been established for a long time and has experienced a modest growth rate over the last 20 years. *Language models* have also been researched for a long time. However, the number of publications on this topic has only experienced significant growth since 2018. Similar differences can be observed when looking at the other popular fields of study. *Representation learning* and *text classification*, while generally widely researched, are partially stagnant in their growth. In contrast, *dialogue systems & conversational agents* and particularly *low-resource NLP,* continue to exhibit high growth rates in the number of studies. Based on the development of the average number of studies on the remaining fields of study, we observe a slightly positive growth overall. However, the majority of fields of study are significantly less researched than the most popular fields of study.

Recent Trends in NLP 🚀



Growth-share matrix of fields of study in NLP. The growth rates and total number of works for each field of study are calculated from the start of 2018 to the end of 2022. Image by author.

The figure above shows the growth-share matrix of fields of study in NLP. We use it to examine current research trends and possible future research directions by analyzing the growth rates and total number of papers related to the various fields of study in NLP between 2018 and 2022. The upper right section of the matrix consists of fields of study that exhibit a high growth rate and simultaneously a large number of papers overall. Given the growing popularity of fields of study in this section, we categorize them as *trending stars*. The lower right section contains fields of study that are very popular but exhibit a low growth rate. Usually, these are fields of study that are essential for NLP but already relatively mature. Hence, we categorize them as *foundational fields of study*. The upper left section of the matrix contains fields of study that exhibit a high growth rate but only very few papers overall. Since the progress of these fields of study is rather promising, but the small number of overall papers renders it difficult to predict their further developments, we categorize them as *rising question marks*. The fields of study in the lower left of the matrix are categorized as *niche fields of study* owing to their low total number of papers and their low growth rates.

The figure shows that *language models* are currently receiving the most attention. Based on the latest developments in this area, this trend is likely to continue and accelerate in the near future. *Text classification*, *machine translation*, and *representation learning* rank among the most popular fields of study, but only show marginal growth. In the long term, they may be replaced by faster-growing fields as the most popular fields of study.

In general, fields of study related to *syntactic text processing* exhibit negligible growth and low popularity overall. Conversely, fields of study concerned with *responsible & trustworthy NLP*, such as *green & sustainable NLP*, *low-resource NLP*, and *ethical NLP*, tend to exhibit a high growth rate and high popularity overall. This trend can also be observed in the case of *structured data in*

NLP, *visual data in NLP*, and *speech & audio in NLP*, all of which are concerned with multimodality. In addition, *natural language interfaces* involving *dialogue systems & conversational agents* and *question answering* are becoming increasingly important in the research community. We conclude that in addition to *language models*, *responsible & trustworthy NLP*, *multimodality*, and *natural language interfaces* are likely to characterize the NLP research landscape in the near future.

Further notable developments can be observed in the area of *reasoning*, specifically with respect to *knowledge graph reasoning* and *numerical reasoning* and in various fields of study related to *text generation*. Although these fields of study are currently still relatively small, they apparently attract more and more interest from the research community and show a clear positive tendency toward growth.

Conclusion 💡

To summarize recent developments and provide an overview of the NLP landscape, we defined a taxonomy of fields of study and analyzed recent research developments.

Our findings show that a large number of fields of study have been studied, including trending fields such as *multimodality*, *responsible & trustworthy NLP*, and *natural language interfaces*. We hope that this article provides a useful overview of the current NLP landscape and can serve as a starting point for a more in-depth exploration of the field.

Sources

Exploring the Landscape of Natural Language Processing Research

As an efficient approach to understand, generate, and process natural language texts, research in natural language…

arxiv.org

NLP | Nlproc | Naturallanguageprocessing | Data Science | AI

Follow

Written by Tim Schopf

273 Followers · Writer for Towards Data Science

PhD candidate @ Technical University of Munich | Passionate about NLP and Knowledge Graphs! Get in touch: https://de.linkedin.com/in/tim-schopf