Seminarios - S14: Natural Language Processing Individual work - Corpus collection

Jenifer Tabita Ciuciu-Kiss

9 June 2024

Introduction

This document presents my individual project for the natural language processing seminar. For this assignment, I chose a theoretical exploration of **corpus collection**. I selected this topic because I plan to create my own corpus for my PhD research, and I believe it is important to be well-informed in the current state-of-the-art in this area before starting to work on my corpus collection project. This document can be viewed here on Overleaf.

1 Introduction

A corpus (5), in the context of Natural Language Processing (NLP), is a large and structured set of texts that are used for developing language models and for linguistic research. The texts in a corpus are often collected from various sources and are typically representative of a particular language or domain. The goal of collecting a corpus is to capture a wide range of language use cases, enabling researchers and developers to train and evaluate NLP models effectively.

Corpora are essential in NLP as they provide the practical data needed to build and validate models. These large datasets (26) enable the development of algorithms that can understand and generate human language. Without corpora, it would be challenging to create models that can accurately perform tasks such as translation, sentiment analysis, and information extraction. Corpora serve as the foundation for training machine learning models, allowing them to learn the nuances and complexities of natural language.

Corpus-based methods (33) involves the use of text corpora to inform the development and refinement of NLP techniques. These methods leverage statistical and machine learning approaches to analyze the text and extract meaningful information. Techniques such as part-of-speech tagging, syntactic parsing, and semantic analysis are often based on insights gained from extensive corpus analysis. By utilizing large corpora, these methods can achieve higher accuracy and more robust performance in various NLP applications.

2 Types of corpora

The study and application of NLP strongly rely on the availability and quality of the corpora. A corpus is not just a collection of texts; it is a structured repository that serves various linguistic and computational purposes. The diversity of the available corpora reflects the diverse nature of language and its usage in different contexts. Understanding the different types of corpora is important for selecting the right resources for specific NLP tasks. This section explores the various types of corpora through examples, highlighting their unique characteristics and applications of each of them.

2.1 General corpora

General corpora (32) are large collections of texts that are not limited to a specific domain or subject. They are designed to be representative of a wide range of language uses and genres. Examples include the British National Corpus (BNC) (25) and the Corpus of Contemporary American English (COCA) (9). These corpora are valuable resources for general linguistic research and for training NLP models that need to perform well across diverse contexts.

2.2 Specialized corpora

Specialized corpora (11) are collections of texts that focus on a specific domain or subject area, such as medicine, law, or finance. These corpora are tailored to capture the unique terminology and linguistic patterns of their respective fields. For instance, the MEDLINE (16) corpus contains a vast amount of biomedical literature, making it an essential resource for developing NLP tools in the medical domain. Specialized corpora enable researchers to build models that are more accurate and effective in specific contexts.

2.3 Multilingual corpora

Multilingual corpora (40) consist of texts in multiple languages and are often used for tasks such as machine translation and cross-lingual information retrieval. Examples include the Europarl corpus (23), which contains proceedings from the European Parliament (20; 19) in multiple languages, and the United Nations Parallel Corpus (44), which includes documents translated into the UN's official languages. Multilingual corpora are important for developing models that can handle multiple languages and for studying linguistic phenomena across different languages.

2.4 Annotated corpora

Annotated corpora (35) are texts that have been enriched with additional linguistic information, such as part-of-speech tags, syntactic trees, and semantic annotations. These corpora are essential for training supervised machine learning models, as they provide the ground truth needed to learn various linguistic tasks. The Penn Treebank (41), for example, is a widely used annotated corpus that includes syntactic annotations for English text. Annotated corpora enable the development of sophisticated NLP systems that can understand and generate human language with high accuracy.

3 Building a corpus

Building a high-quality corpus (37) is an elementary step in many NLP projects because many times there is not available corpora for the specific task or domain. The process involves collecting a diverse and representative set of texts, ensuring that the data is properly annotated, and addressing legal and ethical considerations. This section presents the most important methods of data collection, as well as important legal and ethical issues to consider during the corpus-building process.

3.1 Data collection methods

Effective data collection (2; 29) is crucial for creating a representative and useful corpus. Several methods can be used to gather textual data, each with its own advantages and challenges. This subsection discusses three common data collection methods: web scraping, crowdsourcing, and collaboration with institutions.

3.1.1 Web scraping

Web scraping (30) is a popular method for collecting large volumes of text data from the internet. It involves using automated tools to extract information from websites. This method is particularly useful for gathering up-to-date information and can cover a wide range of topics and genres. However, web scraping requires careful attention to the terms of service of websites and compliance with legal restrictions to avoid potential issues.

3.1.2 Crowdsourcing

Crowdsourcing (22) leverages the contributions of a large number of individual experts in a given domain to collect and annotate text data. Platforms like Amazon Mechanical Turk (42) enable researchers to gather diverse datasets quickly and cost-effectively. Crowdsourcing is especially useful for obtaining annotations for large datasets, as it allows for distributed work among many contributors. Ensuring the quality of crowdsourced data can be challenging, but methods such as consensus scoring and qualification tests can help maintain high standards.

3.1.3 Collaboration with institutions

Collaborating with academic and industry institutions (8) can provide access to specialized datasets that are otherwise difficult to obtain. Institutions often have large collections of domain-specific texts, such as medical records, legal documents, or scientific publications. These collaborations can enhance the quality and relevance of the corpus, particularly for specialized NLP applications. Establishing partnerships with institutions may also involve negotiating data-sharing agreements and ensuring compliance with institutional policies.

3.2 Legal and ethical considerations

Addressing legal and ethical considerations (38) are super important when building a corpus. Ensuring that the data collection process respects copyright laws and protects the privacy of individuals involved is mandatory. This section discusses on a high level the key legal and ethical issues, including copyright compliance and privacy protection.

3.2.1 Copyright issues

When building a corpus, it is essential to address copyright issues (7) to ensure that the data collection process is legal. Texts that are still under copyright protection require proper permissions from the copyright holders. Researchers can use texts that are in the public domain or covered by licenses that allow academic use, such as Creative Commons licenses (17). Violating copyright laws can lead to legal repercussions and undermine the integrity of the research.

3.2.2 Anonymization and privacy concerns

Privacy concerns are paramount when the corpus includes personal or sensitive information. Anonymization techniques (4) must be employed to protect the identities of individuals mentioned in the texts. This can involve removing or obscuring names, addresses, and other identifying details. Ensuring compliance with data protection regulations, such as the General Data Protection Regulation (GDPR) (43) in the European Union, is critical. Ethical guidelines, such as those provided by the Association for Computational Linguistics (ACL) (21), offer frameworks for conducting responsible research in NLP.

4 Corpus annotation

Corpus annotation (24) involves adding metadata and linguistic information to raw text, making it even more useful for NLP tasks. Annotated corpora provide the ground truth needed to train and evaluate models. This section covers the types of annotations commonly used in NLP, the tools available for annotation, and the importance of inter-annotator agreement.

4.1 Types of annotations

Annotations (36) can vary widely depending on the goals of the NLP project. This pat present the main different types of annotations, including Part-of-Speech (POS) tagging (28), Named Entity Recognition (NER) (27), syntactic parsing (12), and semantic roles (14).

4.1.1 Part-of-Speech (POS) tagging

Part-of-Speech (POS) tagging (28) involves labelling each word in a text with its corresponding part of speech, such as noun, verb, adjective, etc. POS tagging is fundamental for many NLP tasks as it helps in understanding the grammatical structure of sentences. Accurate POS tagging is crucial for higher-level tasks like syntactic parsing and semantic analysis.

4.1.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) (27) is the process of identifying and classifying proper nouns in text into predefined categories such as person names, organizations, locations, dates, etc. NER is essential for information extraction tasks, helping systems understand and organize unstructured text data by categorizing entities mentioned within it.

4.1.3 Syntactic parsing

Syntactic parsing (12) involves analyzing the grammatical structure of sentences and representing this structure in the form of a parse tree. Each node in the tree represents a syntactic constituent, such as a noun phrase or verb phrase. Syntactic parsing is critical for understanding the hierarchical nature of language and is used in various applications like machine translation and question answering.

4.1.4 Semantic Roles Labelling (SRL)

Semantic Role Labelling (SRL) (14; 18) is the process of assigning labels to words or phrases in a sentence that indicate their semantic role in the context of the sentence. For example, identifying the agent, patient, and instrument in a sentence. SRL helps in understanding the meaning of sentences by identifying who did what to whom, when, and how.

4.2 Inter-Annotator Agreement (IAA)

Inter-annotator agreement (IAA) (1) is a measure of the consistency between different annotators who independently label the same dataset. It is an essential aspect of corpus annotation as it reflects the reliability and accuracy of the annotations. High IAA indicates that the annotators are applying the annotation guidelines consistently, which is crucial for the quality and validity of the corpus. Conversely, low IAA may suggest ambiguities in the annotation guidelines or differences in the annotators' understanding of the task, necessitating further clarification and training.

IAA is essential for the reliability of annotated corpora. This subsection explores methods for measuring consistency between annotators and strategies for resolving discrepancies.

4.2.1 Measuring consistency

IAA (1) is measured to assess the consistency of annotations provided by different annotators. Common metrics for measuring IAA include Cohen's Kappa, Fleiss' Kappa, and Krippendorff's Alpha. These metrics help in quantifying the degree of agreement and identifying areas where annotators diverge.

4.2.2 Resolving discrepancies

When discrepancies between annotators occur, it is important to resolve them to maintain the quality of the annotated corpus. This can be done through regular calibration sessions where annotators discuss and reconcile their differences. Clear and detailed annotation guidelines also play a crucial role in minimizing discrepancies. In some cases, a third-party adjudicator may be used to make the final decision on disputed annotations.

5 Corpus utilization

Once a corpus has been collected and annotated, it can be used in a variety of ways to advance NLP research and applications. This section explores how corpora are utilized in training machine learning models, conducting linguistic research, and evaluating NLP systems. Generative AI techniques and applications are not included.

5.1 Training Machine Learning (ML) models

Corpora play a crucial role in training ML models for various NLP tasks. This section introduces some methods of using corpora for supervised and unsupervised learning.

5.1.1 Supervised learning

In supervised learning, annotated corpora are used to train models on specific tasks by providing labelled examples. These tasks can range from text classification and sentiment analysis to more complex applications like machine translation and question answering. The annotated data serves as the ground truth that the models learn to predict. High-quality annotated corpora are essential for achieving high accuracy and robust performance in supervised learning tasks.

5.1.2 Unsupervised learning

Unsupervised learning involves using corpora without explicit annotations to discover patterns and structures within the data. Techniques such as clustering and topic modelling are common applications of unsupervised learning in NLP. These methods help in identifying underlying themes and grouping similar texts together based on their content. Unsupervised learning is particularly useful when labelled data is lacking or unavailable.

5.2 Linguistic research

Corpora are invaluable resources for conducting linguistic research (3). They provide practical data that can be analyzed to gain insights into various linguistic phenomena. This subsection discusses the use of corpora in lexical studies and syntactic and semantic analysis.

5.2.1 Lexical studies

Lexical studies (39) involve analyzing the words and phrases used in a language, including their frequency, distribution, and contextual usage. Corpora provide large datasets that enable researchers to study lexical variations and trends across different text genres and time-periods. Such studies can reveal important insights into language change and vocabulary usage patterns.

5.2.2 Syntactic and semantic analysis

Syntactic and semantic analysis (10) aims to understand the structure and meaning of sentences. Syntactic analysis involves studying how words are combined to form grammatical sentences, while semantic analysis focuses on the meanings conveyed by these sentences. Corpora with syntactic and semantic annotations allow researchers to examine these aspects systematically, facilitating the development of theories and models that explain language structure and meaning.

6 Challenges in corpus utilization

While corpora are important resources for NLP, their utilization comes with several challenges. Addressing these challenges is crucial for ensuring the effectiveness and reliability of NLP applications. This section discusses the major challenges related to data quality, scalability, and bias in corpora.

6.1 Data quality

The quality of data in a corpus significantly impacts the performance of NLP models. This part examines the issue of noisy data and its implications.

6.1.1 Noisy data and its impact

Noisy data includes errors such as misspellings, grammatical mistakes, and irrelevant information. These issues can arise from various sources, including web scraping and OCR (optical character recognition) (31) errors. Noisy data can decrease the performance of NLP models by introducing inaccuracies and inconsistencies. It is essential to implement preprocessing steps such as data cleaning and normalization to mitigate the effects of noisy data and improve the overall quality of the corpus.

6.1.2 Improving data quality

Improving data quality involves several strategies, such as employing automated data cleaning tools, manual review, and validation processes. Automated tools can detect and correct common errors like misspellings and grammatical mistakes. Additionally, employing domain experts to manually review and validate the data can ensure higher accuracy and relevance. Regular updates and maintenance of the corpus are also necessary to incorporate the latest language usage and trends, thereby keeping the corpus relevant and high-quality.

6.2 Scalability issues

Handling large volumes of text data presents scalability challenges. This subsection explores strategies for managing and processing large corpora effectively.

6.2.1 Handling large corpora

As the size of a corpus increases, so do the computational resources required to store, process, and analyze the data. Efficient data storage solutions, such as distributed databases and cloud storage, are essential for managing large corpora. Additionally, parallel processing and optimization techniques can help in handling large-scale data processing tasks. Implementing scalable algorithms and leveraging high-performance computing resources can ensure that NLP applications remain efficient and responsive even with extensive corpora.

6.2.2 Optimizing computational resources

Optimizing computational resources involves using advanced hardware and software solutions to manage large datasets. Techniques such as data partitioning (34), distributed computing (6), and in-memory processing (15) can significantly reduce the time and resources required for data analysis. Using cloud-based platforms (13) can provide scalable and flexible resources that adjust to the workload demands. Moreover, employing efficient algorithms and data structures can improve processing speeds and reduce computational overhead.

6.3 Bias in corpora

Bias in corpora can lead to biased NLP models, which in turn can produce unfair or inaccurate outcomes. This subsection discusses the sources of bias and strategies for mitigation.

6.3.1 Sources of bias

Bias in corpora can stem from various sources, including the selection of texts, the demographics of the authors, and the contexts in which the texts were produced. For example, if a corpus predominantly consists of texts from a particular region or social group, the resulting models may not generalize well to other regions or groups. Additionally, historical biases and stereotypes present in the data can be perpetuated by NLP models, leading to biased outputs.

6.3.2 Mitigation strategies

Mitigating bias in corpora involves several strategies. One approach is to ensure diversity and representativeness in the selection of texts. This can involve including texts from various regions, social groups, and genres to create a more balanced corpus. Another strategy is to apply bias detection and correction techniques during data preprocessing and model training. Regular audits and evaluations of the corpus and the models trained on it can help identify and address biases, ensuring more fair and accurate NLP applications.

7 Conclusion

In conclusion, the process of corpus collection and utilization in NLP is a complex field that encompasses various stages, from data collection to annotation and utilization. This document has provided an overview of the types of corpora, methods for building them, and the critical aspects of their annotation and utilization.

Understanding the different types of corpora is important for selecting the right resources for specific NLP tasks. General corpora provide a broad base for linguistic research and model training, while specialized corpora cater to domain-specific applications. Multilingual corpora are invaluable for tasks involving multiple languages, and annotated corpora provide the necessary labelled data for supervised learning models.

Building a high-quality corpus involves effective data collection methods such as web scraping, crowdsourcing, and collaboration with institutions. It also requires careful attention to legal and ethical considerations, including copyright issues and privacy concerns.

Annotation is a crucial step that enriches raw text with metadata and linguistic information, making it more useful for NLP tasks. Different types of annotations, such as POS tagging, NER, syntactic parsing, and semantic roles, serve various purposes in linguistic research and model training. Ensuring high inter-annotator agreement is vital for the reliability of annotated corpora.

Corpora are used in training ML models, conducting linguistic research, and evaluating NLP systems. Supervised and unsupervised learning techniques leverage these corpora to develop models that can understand and generate human language. Corpora also provide the empirical data needed for lexical studies and syntactic and semantic analysis, offering insights into language use and structure.

Despite their importance, using corpora comes with challenges such as data quality, scalability, and bias. Addressing noisy data, managing large corpora, and mitigating bias are essential steps to ensure the effectiveness and fairness of NLP applications.

As the field of NLP continues to evolve, the role of corpora remains central to advancing research and developing robust models. By staying informed about the latest developments in corpus collection and annotation, researchers and practitioners can build better resources and create more accurate and reliable NLP systems.

References

[1] Artstein, R.: Inter-annotator agreement. Handbook of linguistic annotation pp. 297–313 (2017)

[2] Bar-Ilan, J.: Data collection methods on the web for infometric purposes—a review and analysis. Scientometrics **50**, 7–32 (2001)

[3] Bates, E., Wulfeck, B., MacWhinney, B.: Cross-linguistic research in aphasia: An overview. Brain and language **41**(2), 123–148 (1991)

[4] Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: 21st International conference on data engineering (ICDE'05). pp. 217–228. IEEE (2005)

[5] Biber, D., Conrad, S., Reppen, R.: Corpus linguistics: Investigating language structure and use. Cambridge University Press (1998)

[6] Birman, K.P.: The process group approach to reliable distributed computing. Communications of the ACM **36**(12), 37–53 (1993)

[7] Cantrell, M.A., Lupinacci, P.: Methodological issues in online data collection. Journal of advanced nursing **60**(5), 544–549 (2007)

[8] Chompalov, I., Genuth, J., Shrum, W.: The organization of scientific collaborations. Research policy **31**(5), 749–767 (2002)

[9] Davies, M.: The corpus of contemporary american english as the first reliable monitor corpus of english. Literary and linguistic computing **25**(4), 447–464 (2010)

[10] Ferreira, R., Lins, R.D., Simske, S.J., Freitas, F., Riss, M.: Assessing sentence similarity through lexical, syntactic and semantic analysis. Computer Speech & Language **39**, 1–28 (2016)

[11] Flowerdew, L.: The argument for using english specialized corpora to understand academic and professional language. Discourse in the professions: Perspectives from corpus linguistics **11**, 11–33 (2004)

[12] Frazier, L.: On comprehending sentences: Syntactic parsing strategies. University of Connecticut (1979)

[13] Fylaktopoulos, G., Goumas, G., Skolarikis, M., Sotiropoulos, A., Maglogiannis, I.: An overview of platforms for cloud based development. SpringerPlus **5**, 1–13 (2016)

[14] Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational linguistics **28**(3), 245–288 (2002)

[15] Grady, C.L., Craik, F.I.: Changes in memory processing with age. Current opinion in neurobiology **10**(2), 224–231 (2000)

[16] Greenhalgh, T.: How to read a paper: the medline database. Bmj **315**(7101), 180–183 (1997)

[17] Hagedorn, G., Mietchen, D., Morris, R.A., Agosti, D., Penev, L., Berendsohn, W.G., Hobern, D.: Creative commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. ZooKeys (150), 127 (2011)

[18] He, L., Lee, K., Lewis, M., Zettlemoyer, L.: Deep semantic role labeling: What works and what's next. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 473–483 (2017)

[19] Jacobs, F.: The european parliament. In: Reforming the European Union, pp. 57–73. Routledge (2014)

[20] Judge, D., Earnshaw, D.: The European Parliament (2003)

[21] Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.: Proceedings of the 58th annual meeting of the association for computational linguistics. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)

[22] Karger, D.R., Oh, S., Shah, D.: Efficient crowdsourcing for multi-class labeling. In: Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems. pp. 81–92 (2013)

[23] Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of machine translation summit x: papers. pp. 79–86 (2005)

[24] Leech, G.: Corpus annotation schemes. Literary and linguistic computing **8**(4), 275–281 (1993)

[25] Leech, G.N.: 100 million words of english: the british national corpus (bnc). (1992)

[26] Manning, C., Schutze, H.: Foundations of statistical natural language processing. MIT press (1999)

[27] Mansouri, A., Affendey, L.S., Mamat, A.: Named entity recognition approaches. International Journal of Computer Science and Network Security **8**(2), 339–344 (2008)

[28] Martinez, A.R.: Part-of-speech tagging. Wiley Interdisciplinary Reviews: Computational Statistics **4**(1), 107–113 (2012)

[29] Mazhar, S.A., Anjum, R., Anwar, A.I., Khan, A.A.: Methods of data collection: A fundamental tool of research. Journal of Integrated Community Health (ISSN 2319-9113) **10**(1), 6–10 (2021)

[30] Mitchell, R.: Web scraping with Python: Collecting more data from the modern web. " O'Reilly Media, Inc." (2018)

[31] Mithe, R., Indalkar, S., Divekar, N.: Optical character recognition. International journal of recent technology and engineering (IJRTE) **2**(1), 72–75 (2013)

[32] Nelson, M.: Building a written corpus. The Routledge handbook of corpus linguistics pp. 53–65 (2010)

[33] Parsing, C.: Speech and language processing. Power Point Slides (2009)

[34] Picard, R.R., Berk, K.N.: Data splitting. The American Statistician **44**(2), 140–147 (1990)

[35] Poesio, M., Pradhan, S., Recasens, M., Rodriguez, K., Versley, Y.: Annotated corpora and annotation tools. Anaphora Resolution: Algorithms, Resources, and Applications pp. 97–140 (2016)

[36] Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media, Inc." (2012)

[37] Reppen, R.: Building a corpus: what are key considerations? In: The Routledge handbook of corpus linguistics, pp. 13–20. Routledge (2022)

[38] Safdar, N.M., Banja, J.D., Meltzer, C.C.: Ethical considerations in artificial intelligence. European journal of radiology **122**, 108768 (2020)

[39] Saucier, G., Goldberg, L.R.: Lexical studies of indigenous personality factors: Premises, products, and prospects. Journal of personality **69**(6), 847–879 (2001)

[40] Schmidt, T., Wörner, K.: Multilingual corpora and multilingual corpus analysis, vol. 14. John Benjamins Publishing (2012)

[41] Taylor, A., Marcus, M., Santorini, B.: The penn treebank: an overview. Treebanks: Building and using parsed corpora pp. 5–22 (2003)

[42] Turk, A.M.: Amazon mechanical turk. Retrieved August **17**, 2012 (2012)

[43] Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)

[44] Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B.: The united nations parallel corpus v1. 0. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 3530–3534 (2016)