16-06-2023

**INTRODUCTION TO NLP- UNIVERSITY OF LONDON (2)**

**Chapter 3: Getting to Grips with Natural language Data:**

1. **Corpora and Other Data Resources:**

A corpus is a large collection of linguistic data, which can consist of written texts, transcriptions of speech, or a combination of recorded speech and transcriptions. Corpora are essential in NLP because they provide the data needed to analyse and model language patterns.

Three broad categories of corpora:

**1. Monolingual corpora:** These corpora contain data from a single language. They are used to study and analyse specific linguistic properties and patterns within that language.

**2. Comparable corpora:** Comparable corpora consist of multiple monolingual corpora in different languages. Ideally, these corpora have a similar level of balance and representativeness. Comparable corpora allow for contrastive studies between languages, comparing and analysing their similarities and differences.

**3. Parallel corpora:** Parallel corpora include original texts in one language and their translations in one or more different languages. These corpora are valuable for training statistical translation systems and developing machine translation algorithms.

Corpora are generally expected to have additional characteristics. They are often constructed to be balanced and representative of a particular domain. For quantitative analysis, it is crucial to ensure that the corpus is not skewed or unrepresentative, as this could lead to unreliable results. However, some corpora may be compiled without strict systematic construction, such as "the web as corpus."

Corpus builders need to exercise expert judgment when determining the sampling frame, which represents the population of texts from which the samples are taken. The size of the corpus is also a consideration for maximal representativeness. The sampling frame can be bibliographic, based on comprehensive indexes or library holdings, or demographic, selecting informants based on social categories, similar to public opinion research.

Corpora have traditionally been of finite size and remain fixed once compiled. However, there are also monitor corpora that are continually updated with new material. **Monitor corpora** are especially useful for tracking new words entering the language and monitoring changes in word usage over time. An example of a monitor corpus is the COBUILD Bank of EnglishTM, which has grown from around 300 million words to 650 million words.

Corpora can be further classified based on their structure.

1. They can be **isolated**, consisting of an unorganized collection of individual texts, or categorized, where texts are organized by categories such as genre.
2. **Overlapping corpora** contain categories that overlap, such as a news corpus covering both politics and sports.
3. **Temporal corpora** indicate language use over time, allowing for the study of language changes. Examples include the Inaugural corpus and the Helsinki Diachronic corpus.

There are also several specific corpora, such as the Brown corpus, BNC (British National Corpus), COBUILD (Bank of English), Gutenberg corpus, and Penn Treebank. These corpora are widely used in linguistic research and NLP applications.

Overall, corpora are valuable resources in natural language processing, providing the data needed for language analysis, modeling, translation, and other language-related tasks.

1. **Some Uses of Corpora:**
2. Lexicography:

the focus on lexicography refers to the practice of compiling dictionaries. Modern dictionaries, such as Chambers, Collins, and Longmans, heavily rely on corpus data to classify and document the various ways words are used in contemporary English. By utilizing corpus data, lexicographers can accurately capture the different meanings and usages of words, as well as any changes that may have occurred over time.

For example, lexicographers can use corpus data to determine whether certain words can be used as verbs. By entering the appropriate search query, they can retrieve examples from the corpus that demonstrate the verb usage of words like "scapegoat," "thermostat," or "leverage." These examples can help lexicographers understand how these words are used in context. Here are some sample sentences provided from the British National Corpus (BNC) for illustration:

1. Scapegoating an individual prevents the debate and delays community understanding.

2. The measuring cell is immersed in a vat of liquid, usually benzene or xylene, which can be thermos tatted at temperatures between 273 and 400 K.

3. These one-time costs, once met, could be leveraged over much more business activity around the globe than we then enjoyed.

In summary, corpus data plays a crucial role in modern lexicography by providing lexicographers with real-world examples of word usage, enabling them to accurately document and classify the various ways words are used in contemporary language.

1. Grammar and Syntax:

Grammars such as the Comprehensive Grammar of the English Language and the Cambridge Grammar of the English Language utilize corpora as one of their sources of information, in addition to linguistic research findings and the compilers' own intuitive knowledge as competent speakers of the language.

Corpora provide valuable data that can be used to support and inform grammatical descriptions. In the past, the use of corpora in grammar development primarily involved qualitative analysis, where linguists would examine the data qualitatively to identify patterns and make observations. However, recent advancements in computational power and the availability of parsed corpora and analysis tools have enabled researchers to conduct quantitative studies as well.

Quantitative studies using corpora can involve analysing various aspects of grammatical frequency, such as the relative frequency of different clause types in English. By examining a large amount of language data, researchers can gain insights into how frequently certain grammatical constructions or patterns occur in real-world usage.

Corpora also play a role in testing the predictions made by formal grammars developed within the generative school of linguistics. Generative grammars propose formal rules and structures for generating sentences. By analysing corpora, linguists can investigate whether the predictions made by these grammars align with the patterns observed in actual language usage.

The COBUILD project, which provided resources for Collins English dictionaries, has resulted in a series of small handbooks covering various grammatical constructions. These resources are designed to be useful for both advanced language learners and linguists seeking examples and insights into different grammatical phenomena.

In summary, corpora are an important tool in the development of large-scale grammars for pedagogic and reference use. They provide empirical data that can support linguistic research, help identify grammatical patterns and frequencies, and test the predictions of formal grammars. Corpora contribute to a more accurate and comprehensive understanding of the grammar of a language and are valuable resources for language learners and linguists alike.

1. Stylistics: Variation across authors, periods, genres and channels of communication:

Style refers to the choices individuals make when expressing themselves, and it suggests that people have various options for how they communicate. These choices may not only be consistent for individuals across different instances of communication but also reflect certain characteristics of genres, time periods, and channels of communication.

When we refer to genres, we mean different types of communication, such as romantic fiction, financial news, court reports, and more. Each genre has its own conventions and preferred ways of expression. For example, romantic fiction tends to use poetic language and vivid descriptions, while financial news relies on concise and factual language.

Channels of communication refer to different modes of transmitting information, such as written text and spoken discourse. Each channel can further be divided into specific subtypes. For instance, people may make different choices when composing emails, sending text messages, or writing a letter by hand. Similarly, individuals may adopt different styles when communicating face-to-face compared to when they are on the telephone.

The notion of style becomes relevant in various contexts. Literary scholars, as well as law enforcement and intelligence agencies, may be interested in identifying the author of a document based on internal evidence. They examine the specific choices made in language use, such as vocabulary, grammar, and writing style, to make inferences about the potential author. This process is known **as attribution of authorship**. In some cases, there have been debates and controversies regarding the true authors of famous works, such as Shakespeare's plays, where various figures have been proposed as potential authors.

1. Training and Evaluation:

Corpora play a crucial role in various tasks, including:

**1. Detecting the topic of a document:** Corpora are used to train machine-learning systems to identify the subject or main theme of a given text.

**2. Sentiment analysis:** Corpora enable the development of machine-learning models to analyse and determine the sentiments expressed in a text, whether they are positive, negative, or neutral towards a particular product or policy.

**3. Named entity recognition**: Corpora help in training machine-learning systems to identify and extract specific individuals mentioned in a text, as well as the relationships between them and the events they are involved in.

**4. Statistical parsing:** Corpora are utilized for training machine-learning models to perform syntactic parsing, which involves analyzing the grammatical structure of sentences and identifying their constituents.

**5. Statistical machine translation:** Corpora serve as training data for machine translation systems, enabling them to learn patterns and align corresponding sentences in different languages.

For example, the Brown corpus and the Wall Street Journal (WSJ) corpus are commonly used to evaluate text segmentation and other text processing tasks. These corpora provide a benchmark for assessing the performance of machine-learning systems in specific applications.

These tasks typically involve supervised learning, where machine-learning systems are trained on corpora that have been manually annotated by human annotators. The corpus is divided into training and test sets, with the test set serving as a gold standard for comparing the accuracy of the trained system with that of the annotators. Multiple annotators are often employed to assess inter-annotator agreement, which sets an upper bound for the system's performance.

It is important to note that human annotators may not always agree, and it is unrealistic to expect a computer program to achieve 100% accuracy on tasks where human annotators themselves have disagreements. Nevertheless, corpora provide valuable training data for machine-learning systems and serve as a basis for evaluating their performance.

1. **Corpora:**
2. Brown Corpus:

the Brown corpus holds historical significance as one of the pioneering large-scale corpora in the field of computational linguistics. It offers valuable resources for studying and analysing the characteristics of written texts in different genres and serves as a foundation for language research and development of language technologies.

Brown corpus, which was one of the first large-scale machine-readable corpora. While it may seem relatively small compared to modern standards, consisting of one million words, it was a significant development in its time. The corpus was created at Brown University and took approximately two decades to complete, starting from the early 1960s.

The Brown corpus aimed to be a standard collection of present-day edited American English. It is organized by genre, with specific headings representing different types of texts.

The Brown corpus is available in both tagged and untagged versions and can be accessed using various methods as described by Bird et al. (2009, Table 2.3). The Natural Language Toolkit (NLTK), a popular library for working with human language data in Python, includes the Brown corpus and provides convenient access to it.

1. British National Corpus:

The British National Corpus (BNC) is a large corpus created and managed by the BNC consortium, which includes Oxford and Lancaster universities, dictionary publishers OUP, Longmans, and Chambers, and the British Library. It was developed from 1991 to 1994 and contains 100 million words, with 90% of the data being written texts and 10% transcriptions of speech. Notably, the BNC was one of the first corpora to include spontaneous spoken English.

The corpus was marked up using an automated part-of-speech tagger, which helped save time and expenses compared to manual annotation by language experts. However, this automated process introduces a certain degree of error, which users may encounter during their analysis.

The BNC can be accessed online, and users can perform various types of analysis using the Simple Query language, allowing them to explore and investigate the corpus for research and linguistic studies.

1. CoBuild Bank of English:

The COBUILD project was a collaboration between Collins Dictionaries and the University of Birmingham. As part of this project, the Collins corpus was developed, which is a vast analytical database of English comprising 2.5 billion words. The corpus includes written material sourced from websites, newspapers, magazines, and books published worldwide, as well as spoken material from radio, TV, and everyday conversations.

Within the Collins Corpus, there is a subsection called the Bank of EnglishTM, which contains 650 million words. The Bank of EnglishTM was utilized as a foundational resource for creating the Collins Advanced Learner's Dictionary, grammar materials, and various tutorials targeted towards English language learners.

While the Collins corpus is not included in the NLTK (Natural Language Toolkit), there is limited online access available to explore and utilize this extensive linguistic resource.

1. Penn Treebank:

The Penn Treebank is a widely used linguistic resource that is frequently employed by empirical researchers. It consists of a collection of texts in various formats. The texts are available in their original raw form, as well as tagged with part-of-speech (POS) labels using a tag set specifically developed for the project. Additionally, the texts are parsed and annotated with constituent structure.

The Penn Treebank project originated at the University of Pennsylvania in the 1990s. It has produced treebanks, which are annotated collections of linguistic data, from several corpora including Brown, Switchboard, ATIS, and the Wall Street Journal (WSJ) in English. Treebanks for Arabic and Chinese languages have also been created as part of this project.

The results of the Penn Treebank project have served as a foundation for further annotation efforts, focusing on areas such as semantics and rhetorical structure. The NLTK (Natural Language Toolkit) includes a selection from the Wall Street Journal component of the Treebank. Researchers can access this subset in different formats, including raw text, POS-tagged, parsed, and combined formats. Additionally, a simplified POS tag set is available for the Wall Street Journal portion of the Treebank within the NLTK.

1. Gutenberg Archive:

The NLTK (Natural Language Toolkit) incorporates a limited collection of literary texts from Project Gutenberg. Project Gutenberg is an online archive that offers free electronic books. The NLTK's selection consists of texts that are no longer under copyright protection, making them available for public use.

Among the literary texts included in the NLTK from Project Gutenberg are works by various authors. These include novels by Jane Austen, such as "Emma" and "Persuasion." The collection also features stories from GK Chesterton's "Father Brown" series and his novel "The Man Who Was Thursday." Additionally, poems by William Blake, John Milton's epic poem "Paradise Lost," and several plays by William Shakespeare, such as "Julius Caesar," "Macbeth," and "Hamlet," are part of the NLTK's selection.

1. Other Corpora:

The NLTK (Natural Language Toolkit) includes several additional corpora that are accessible to users. Some of these corpora are:

**1. Reuters Corpus:** This corpus consists of 10,788 news documents from the Reuters news agency. It contains a total of 1.3 million words and is divided into "training" and "test" sets. The partitioning is designed for training and testing machine learning algorithms, which is discussed further in Chapter 5 of the subject guide.

**2. US Presidents' Inaugural and State of the Union Addresses:** This corpus includes separate files for the inaugural addresses and State of the Union speeches delivered by U.S. Presidents.

**3. UN Declaration of Human Rights:** This corpus contains the Universal Declaration of Human Rights translated into over 300 languages. It includes excerpts such as "All human beings are born free and equal in dignity and rights" translated into different languages.

Additionally, the NLTK provides access to other corpora through online query interfaces, including:

1. **The Corpus of Contemporary American English (COCA):** Hosted at Brigham Young University, this corpus is known as the only large and balanced corpus of American English.
2. **The Intellitext project at the University of Leeds:** This project aims to facilitate corpus use for academics in various humanities fields. It offers access to monolingual and parallel corpora in several European and Asian languages.

These corpora offer researchers and language enthusiasts a wide range of linguistic resources for analysis, language processing tasks, and academic studies.

1. **Some basic corpus analysis:**

Simple techniques for extracting and presenting data in formats that can be easily interpreted by humans.

1. **Concordancing** is a technique that involves identifying and displaying every occurrence of a specific word or phrase within a text or corpus, along with the surrounding context. This allows researchers to examine how the word or phrase is used in different instances.
2. **Collocations** are pairs of words or sequences of words that frequently appear together in a text, suggesting a connection or pattern. Identifying collocations can provide insights into the content or style of a document.
3. **Conditional frequency distributions** enable basic statistical analysis. They involve counting the frequency of events and pairing each event with a specific condition. Conditional frequency distributions can be used to compare word usage across different genres or to analyse word lengths in different languages.

These techniques provide valuable insights into the characteristics and patterns of language use, allowing researchers to make meaningful observations and draw conclusions about various aspects of texts or corpora.

Frequency Distributions:

This example demonstrates a rudimentary stylistic analysis by ranking the Part-of-Speech (POS) tags in the Brown corpus according to their frequency. The breakdown of the steps involved:

1. The Brown corpus is imported from the NLTK library.

2. The different categories within the corpus are listed using the `categories () ` method. In this case, the "science\_fiction" category is of interest.

3. The number of sentences in the "science\_fiction" category is counted using the `len () ` function on the `sents()` method.

4. All word tokens from the "science\_fiction" category, along with their simplified tags, are extracted and stored in the variable `bsf` using the `tagged\_words () ` method with the `simplify\_tags=True` argument.

5. A frequency distribution of the tags is calculated using the `FreqDist () ` function from NLTK, which provides an ordered list of tags paired with their frequency. The resulting frequency distribution is stored in the variable `sf\_tag\_fd`.

6. Finally, the keys (tags) and their corresponding frequencies are displayed using the `keys()` and `tabulate()` methods of the frequency distribution object.

The output shows the 12 most frequent tags in the "science\_fiction" category, along with their respective frequencies. For example, the tag "N" (noun) appears 2232 times, "V" (verb) appears 1473 times, "DET" (determiner) appears 1345 times, and so on.

This example provides a basic overview of the POS tag frequencies in the "science\_fiction" category of the Brown corpus, which can give insights into the linguistic characteristics and style of science fiction texts.