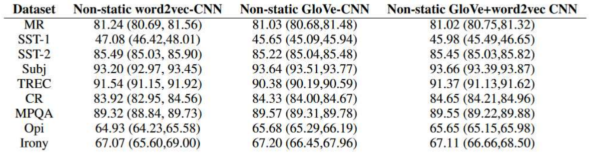
devopedia.org

Text Corpus for NLP

arvindpdmn

25-31 minutos

Summary



Dataset	Non-static word2vec-CNN	Non-static GloVe-CNN	Non-static GloVe+word2vec CNN
MR	81.24 (80.69, 81.56)	81.03 (80.68,81.48)	81.02 (80.75,81.32)
SST-1	47.08 (46.42,48.01)	45.65 (45.09,45.94)	45.98 (45.49,46.65)
SST-2	85.49 (85.03, 85.90)	85.22 (85.04,85.48)	85.45 (85.03,85.82)
Subj	93.20 (92.97, 93.45)	93.64 (93.51,93.77)	93.66 (93.39,93.87)
TREC	91.54 (91.15, 91.92)	90.38 (90.19,90.59)	91.37 (91.13,91.62)
CR	83.92 (82.95, 84.56)	84.33 (84.00,84.67)	84.65 (84.21,84.96)
MPQA	89.32 (88.84, 89.73)	89.57 (89.31,89.78)	89.55 (89.22,89.88)
Opi	64.93 (64.23,65.58)	65.68 (65.29,66.19)	65.65 (65.15,65.98)
Irony	67.07 (65.60,69.00)	67.20 (66.45,67.96)	67.11 (66.66,68.50)

Datasets can help benchmark a model's performance. Source: Zhang and Wallace 2017, table 2.

In the domain of natural language processing (NLP), statistical NLP in particular, there's a need to train the model or algorithm with lots of data. For this purpose, researchers have assembled many text corpora. A common corpus is also useful for benchmarking models.

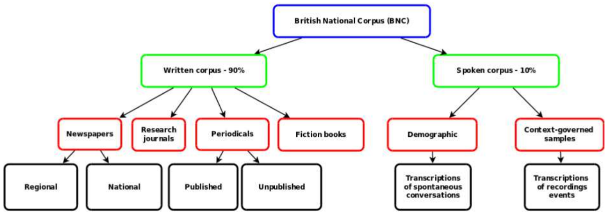
Typically, each text corpus is a collection of text sources. There are dozens of such corpora for a variety of NLP tasks. This article ignores speech corpora and considers only those in text form.

While English has many corpora, other natural languages too have their own corpora, though not as extensive as those for English. Using modern techniques, it's possible to apply NLP on low-resource languages, that is, languages with limited text corpora.

Milestones

W. Nelson Francis and Henry Kučera at the Department of Linguistics, Brown University, publish a computer-readable general corpus to aid linguistic research on modern English. The corpus has 1 million words (500 samples of about 2000 words each). Revised editions appear later in 1971 and 1979. Called **Brown Corpus**, it inspires many other text corpora. The corpus with annotations is included in Treebank-3 (1999).

Linguistic Data Consortium (LDC) is formed to serve as a repository for NLP resources, including corpora. It's hosted at the University of Pennsylvania.



A 100-million corpus of British English called **BNC (British National Corpus)** is assembled between 1991 and 1994. It's balanced across genres. A follow-up task called **BNC2014** is started in 2014, which can help in understanding how language evolves. Spoken BNC2014 is released in September 2017. Written BNC2014 is expected to come out in 2019.

Penn Treebank-3 is released. It's based upon the original Treebank (1992) and its revised Treebank II (1995). This work started in 1989 at the University of Pennsylvania. Treebank-3 includes tagged/parsed Brown Corpus, 1 million words of 1989 WSJ material annotated in Treebank II style, tagged sample of ATIS-3, and tagged/parsed Switchboard Corpus. Apart from POS tags, the corpus includes chunk tags, relation tags and anchor tags. The **BLLIP 1987-89 WSJ Corpus Release 1** has 30 million words and supplements the WSJ section of Treebank-3.

Collected for the years 1990-2007, the **Corpus of Contemporary American English (COCA)** is released with 365 million words. By December 2017, it has 560 million words, adding 20 million each year. There's good balance of spoken, fiction, popular magazines, newspapers, and academic texts. It's been noted that COCA contains many common words that are missing in the American National Corpus (ANC), a corpus of 22 million words.

English Gigaword Fifth Edition is released by LDC. It comes from seven English newswire services. It has 4 billion words and takes up 26 gigabytes uncompressed. The first edition appeared in 2003. In November 2012, researchers at the John Hopkins University add syntactic and discourse structure annotations to this corpus after parsing more than 183 million sentences.

From digitized books, Google releases version 2 of **Google Books Ngrams**. Version 1 came out in July 2009. Only n-grams that appear more than 40 times are included. The corpus includes 1-gram to 5-grams. It includes many non-English languages as well. To experiment on small sets of phrases, researchers can try out the online Google Books Ngram Viewer.

As a corpus for informal genre, **English Web Treebank (EWT)** is released by LDC. This includes content from weblogs, reviews, question-answers, newsgroups, and email. It has about 250K word-level tokens and 16K sentence-level tokens. It's annotated for POS and syntactic structure. This includes Enron Corporation emails from 1999-2002. In 2014, Silveira et al. provide annotation of syntactic dependencies for this corpus that can be used to train dependency parsers.

Common Crawl publishes 240 TiB of uncompressed data from 2.55 billion web pages. Of these, 1 billion URLs were not present in previous crawls. Common Crawl started in 2008. In 2013, they moved from ARC to Web ARChive (WARC) file format. WAT files contain the metadata. WET file contain plaintext of the WARC files.

Discussion

- What are the traits of a good text corpus or wordlist?
  It's said that a prototypical corpus must be machine-readable in Unicode. It must be a representative sample of the language in current use, balanced, and collected in natural settings.

  A good corpus or wordlist must have the following traits:

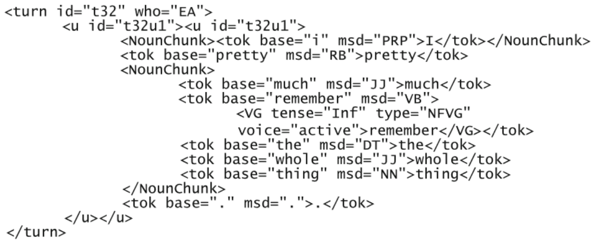
- **Depth**: A wordlist, for instance, should include the top 60K words and not just the top 3K words.
- **Recent**: Corpus based on outdated texts is not going to suit today's tasks.
- **Metadata**: Metadata should indicate the sources, assumptions, limitations and what's included in the corpus.
- **Genre**: Unless corpus has been collected for specific tasks, it should include different genres such as newspapers, magazines, blogs, academic journals, etc.
- **Size**: A corpus of half a million words or more ensures that low frequency words are also adequately represented.

- **Clean**: A wordlist giving word forms of the same word can be messy to process. A better corpus would include only the lemma and part of speech.

- What are the different types of text corpora for NLP?
  A plain text corpus is suitable for **unsupervised** training. Machine learning models learn from the data in an unsupervised manner. However, a corpus that has the raw text plus annotations can be used for **supervised** training. It takes considerable effort to create an annotated corpus but it may produce better results.

  A corpus can be assembled from a variety of sources and genres. Such a corpus can be used for **general** NLP tasks. On the other hand, a corpus might be from a single source, domain or genre. Such a corpus can be used only for a **specific** purpose.

- What are the types of annotations that we can have on a text corpus?



```
<turn id="t32" who="EA">
      <u id="t32u1"><u id="t32u1">
            <NounChunk><tok base="i" msd="PRP">I</tok></NounChunk>
            <tok base="pretty" msd="RB">pretty</tok>
            <NounChunk>
                  <tok base="much" msd="JJ">much</tok>
                  <tok base="remember" msd="VB">
                        <VG tense="Inf" type="NFVG"
                        voice="active">remember</VG></tok>
                  <tok base="the" msd="DT">the</tok>
                  <tok base="whole" msd="JJ">whole</tok>
                  <tok base="thing" msd="NN">thing</tok>
            </NounChunk>
            <tok base="." msd=".">.</tok>
      </u></u>
</turn>
```

  American National Corpus Open annotated with POS, lemma and noun chunks, in XML and standalone form. Source: Gries and Berez 2017, fig. 6.

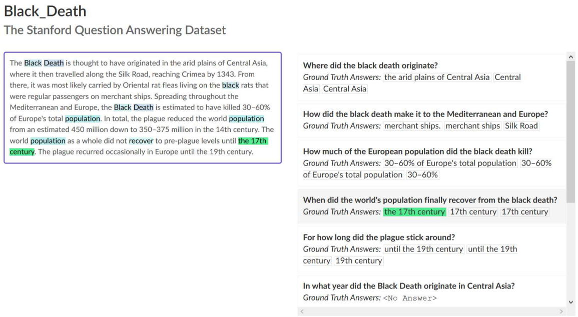
  Part-of-speech is one of the most common annotations because of its use in many downstream NLP tasks. Annotating with lemmas (base forms), syntactic parse trees (phrase-structure or dependency tree representations) and semantic information (word sense disambiguation) are also common. For discourse or text summarization tasks, annotations aid coreference resolutions.

  For instance, British Component of the International Corpus of English (ICE-GB) of 1 million words is POS tagged and syntactically parsed. Another parsed corpus in Penn Treebank. While WordNet and FrameNet are not corpora, they contain useful semantic information.

  Audio/video recordings are transcribed and annotated as well. Annotations are phonetic (sounds), prosodic (variations), or interactional. Video transcripts may annotate for sign language and gesture.

  Annotations could be **inline**/**embedded** with the text. When they appear on separate lines, it's called **multi-tiered** annotation. If they're in separate files, and linked to the text via hypertext, it's called **standalone** annotation.

- What are some NLP task-specific training corpora?



Black_Death

The Stanford Question Answering Dataset

The Black Death is thought to have originated in the arid plains of Central Asia, where it then travelled along the Silk Road, reaching Crimea by 1343. From there, it was most likely carried by Oriental rat fleas living on the black rats that were regular passengers on merchant ships. Spreading throughout the Mediterranean and Europe, the Black Death is estimated to have killed 30–60% of Europe's total population. In total, the plague reduced the world population from an estimated 450 million down to 350–375 million in the 14th century. The world population as a whole did not recover to pre-plague levels until the 17th century. The plague recurred occasionally in Europe until the 19th century.

Where did the black death originate?
Ground Truth Answers: the arid plains of Central Asia | Central Asia | Central Asia

How did the black death make it to the Mediterranean and Europe?
Ground Truth Answers: merchant ships. | merchant ships | Silk Road

How much of the European population did the black death kill?
Ground Truth Answers: 30–60% of Europe's total population | 30–60% of Europe's total population | 30–60%

When did the world's population finally recover from the black death?
Ground Truth Answers: the 17th century | 17th century | 17th century

For how long did the plague stick around?
Ground Truth Answers: until the 19th century | until the 19th century | 19th century

In what year did the Black Death originate in Central Asia?
Ground Truth Answers: <No Answer>

  Example questions and answers from SQuAD. Source: SQuAD

2019b.

Here are some task-specific corpora:

- **POS Tagging**: Penn Treebank's WSJ section is tagged with a 45-tag tagset. Use Ritter dataset for social media content.
- **Named Entity Recognition**: CoNLL 2003 NER task is newswire content from Reuters RCV1 corpus. It considers four entity types. WNUT 2017 Emerging Entities task and OntoNotes 5.0 are other datasets.
- **Constituency Parsing**: Penn Treebank's WSJ section has dataset for this purpose.
- **Semantic role labelling**: OntoNotes v5.0 is useful due to syntactic and semantic annotations.
- **Sentiment Analysis**: IMDb has released 50K movie reviews. Others are Amazon Customer Reviews of 130 million reviews, 6.7 million business reviews from Yelp, and Sentiment140 of 160K tweets.
- **Text Classification/Clustering**: Reuters-21578 is a collection of news documents from 1987 indexed by categories. 20 Newsgroups is another dataset of about 20K documents from 20 newsgroups.
- **Question Answering**: Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset with 100K questions plus 50K unanswerable questions. Jeopardy dataset of about 200K Q&A is another example.

- Could you list some NLP text corpora by genre?
  Formal genre is typically from books and academic journals. Examples are Project Gutenberg EBooks, Google Books Ngrams, and arXiv Bulk Data Access. There are many text corpora from newswire. Examples are 20 Newsgroups and Reuters-21578.

  For **informal** genre, we can include web data and emails. Corpora for these include Common Crawl, Blogger Corpus, Wikipedia Links Data, Enron Emails, and UCI's Spambase. Corpora derived from reviews include Yelp Reviews, Amazon Customer Reviews, and IMDb Movie Reviews. Even more informal are SMS and tweets, for which we have Sentiment140, Twitter US Airline Sentiment, and SMS Spam Collection.

  Spoken language is often different from written language. 2000 HUB5 English is a dataset that's a transcription of 40 telephone conversations. **Signed language** can also be annotated and transcribed to create a corpus.

  Since languages evolve, when analyzing **old text**, our models need to be trained likewise. Examples include DOE Corpus (600s-1150s), and COHA (1810s-2000s).

  Another special case is of **learners** who are likely to express ideas differently. The Open Cambridge Learner Corpus contains 10K student responses of 2.9 million words.

  It's also common to have **domain-specific** corpora. For example, BioCreative and GENIA are for biology.

- What are some generic training corpora for NLP?
  Some of the well-known corpora are Brown Corpus, British National Corpus (BNC), Lancaster-Oslo/Beren Corpus (LOB), International Corpus of English (ICE), Corpus of Contemporary American English (COCA), Google Books Ngram Corpus, Penn Treebank-3, English Gigaword Fifth Edition, and OntoNotes Release 5.0.

  Wikipedia was not made for training NLP models but it can be used. We would need to strip markup. Gensim Python package has `gensim.corpora.wikicorpus.WikiCorpus` class to process Wikipedia data.

Generic corpora are usually suited for **language modelling**, which is useful for other downstream tasks such as machine translation and speech recognition. Researchers have suggested using Project Gutenberg EBooks; Penn Treebank of about a million words pre-processed by Mikolov et al. in 2011; WikiText-2 of more than 2 million words; and WikiText-103. Google's one-billion word corpus provides a useful benchmark.

- Derived from text corpus, which datasets are useful for NLP tasks?
  Wordlists such as list of names or stopwords are useful for NLP work. Phrases in English (PIE) is another resource to explore distribution of words and phrases. It's based on the BNC corpus.

  Tagsets are essential for POS tagging, chunking, dependency parsing or constituency parsing. DKPro Core Tagset Reference is an excellent resource. University of Lancaster maintains a multilingual semantic tagset.

  Treebanks go beyond just POS-tagging a corpus. A treebank is an annotated corpus in which grammatical structure is typically represented as a tree structure. Examples are Penn Treebank and CHRISTINE Corpus. Treebanks are useful for evaluating syntactic parsers or as resources for ML models to optimize linguistic analyzers.

  Word embeddings are real-valued vectors representations of words. These have improved many NLP task including language modelling and semantic analysis. While it's possible to learn embeddings from a large corpus, it's easier to start with downloadable embeddings. Two sources for downloads are Polyglot and Nordic Language Processing Laboratory (NLPL).

  Perhaps by 2020, we'll be able to download pretrained **language models** and apply it to a variety of NLP tasks.

- Which are some corpora for non-English languages?
  For machine translation, it's common to have **parallel corpus**, that is, aligned text in multiple languages. We mention a few examples:

- Aligned Hansards of the 36th Parliament of Canada containing 1.3 million pairs of aligned text segments in English and French
- Europarl parallel corpus from 1996-2011 of 21 European languages from parliament proceedings
- WMT 2014 EN-DE and WMT 2014 EN-FR
- A corpus using Wikipedia across 20 languages, 36 bitexts, about 610 million tokens and 26 million sentence fragments

  An excellent source is OPUS, the open parallel corpus. Lionbridge published a list of parallel corpora in 2019. Martin Weisser maintains a list that links to many non-English corpora.

- Are there curated lists of datasets for NLP work?
  A simple web search will yield plenty of relevant results. Some include download links to the sources. We mention a few that stand out:

- Linguistic Data Consortium has a list of corpora grouped by project
- English Corpora hosts nine large corpora
- The Stanford NLP Group has shared a list of corpora and treebanks
- Registry of Open Data on AWS stores some NLP-specific datasets
- NLP datasets at fast.ai is actually stored on Amazon S3

- Shared by users, data.world lists 30+ NLP datasets
- Shared by users, Kaggle list wordlists, embeddings and text corpora
- Nicolas Iderhoff's list of NLP datasets includes collection dates and dataset sizes
- Sebastian Ruder tracks NLP progress, organized by tasks, with links to external datasets
- Martin Weisser maintains a list of historical and diachronic corpora
- In NLTK Python code, call `nltk.download()` but we can download them separately as well
- From blogs, three separate lists are from Cambridge Spark, Lionbridge and Open Data Science
- Where can I download text corpora for training NLP models? These are the download links for some notable text corpora:
- Brown Corpus
- Corpus of Contemporary American English (COCA)
- Penn Treebank-3 (paid)
- Data dumps of English Wikipedia
- Wikipedia Links Data
- Project Gutenberg EBooks
- Google Books Ngrams via Google or via Amazon S3 bucket
- arXiv Bulk Data Access
- Common Crawl
- DBpedia 3.5.1 Knowledge Base
- Amazon Customer Reviews
- IMDb Reviews
- Google Blogger Corpus
- Jeopardy Question-Answer Dataset
- Yelp Open Dataset
- Enron Email Dataset
- 20 Newsgroups
- Sentiment140
- SMS Spam Collection
- WordNet

References

1. Al-Rfou, Rami. 2019. "Polyglot." Via Google Sites. Accessed 2019-10-28.
2. Ali, Meiryum. 2019. "The Best 25 Datasets for Natural Language Processing." Lionbridge, July 09. Accessed 2019-10-23.
3. Ali, Meiryum. 2019b. "25 Best Parallel Text Datasets for Machine Translation Training." Lionbridge, June 07. Accessed 2019-10-25.
4. BYU. 2019. "The Corpus of Contemporary American English (COCA) and the American National Corpus (ANC)." Brigham Young University. Accessed 2019-10-25.
5. Barba, Paul. 2019. "Machine Learning for Natural Language Processing." Blog, Lexalytics, March 25. Accessed 2019-10-28.
6. Bies, Ann, Justin Mott, Colin Warner, and Seth Kulick. 2012. "English Web Treebank." LDC2000T43, Philadelphia: Linguistic

Data Consortium August 16. Accessed 2019-10-24.

7. BioCreative. 2019. "Biology corpora." Accessed 2019-10-25.
8. Boukkouri, Hicham El. 2018. "Text Classification: The First Step Toward NLP Mastery." Data From The Trenches, on Medium, June 18. Accessed 2019-10-23.
9. Brownlee, Jason. 2017. "Datasets for Natural Language Processing." Machine Learning Mastery, September 27. Updated 2019-08-07. Accessed 2019-10-23.
10. CASS. 2019. "BNC2014." ESRC Centre for Corpus Approaches to Social Science (CASS). Accessed 2019-10-25.
11. CLiPS. 2018. "Penn Treebank II tag set." CLiPS Research Center, June 22. Accessed 2019-09-07.
12. Cambridge Spark. 2018. "50 free Machine Learning datasets: Sentiment Analysis." Cambridge Spark, on Medium, August 29. Accessed 2019-10-23.
13. Charniak, Eugene, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. "BLLIP 1987-89 WSJ Corpus Release 1." LDC2000T43, Philadelphia: Linguistic Data Consortium. Accessed 2019-10-24.
14. Common Crawl. 2019. "So you're ready to get started." Common Crawl. Accessed 2019-10-25.
15. DKPro Core. 2017. "DKPro Core™ Tagset Reference." Version 1.9.0, December 23. Accessed 2019-10-28.
16. English Corpora. 2019. "Corpus of Contemporary American English." Accessed 2019-10-25.
17. English Corpora. 2019b. "Corpus of Historical American English." Accessed 2019-10-25.
18. Evans, David. 2019. "Corpus building and investigation for the Humanities." University of Birmingham. Accessed 2019-10-28.
19. Fletcher, William H. 2011. "Phrases in English Home." June 28. Accessed 2019-10-28.
20. Google Books. 2012. "Ngram Viewer." Google Books. Accessed 2019-10-25.
21. Graff, David, and Christopher Cieri. 2003. "English Gigaword." LDC2003T05, Philadelphia: Linguistic Data Consortium. Accessed 2019-10-24.
22. Gries, Stefan Th. and Andrea L. Berez. 2017. "Linguistic Annotation in/for Corpus Linguistics." Chapter in: N. Ide and J. Pustejovsky (eds.), Handbook of Linguistic Annotation, Springer Science+Business Media Dordrecht, pp. 379-409. Accessed 2019-10-23.
23. Grigaliūnienė, Jonė. 2015. "Corpora in the Classroom." Vilniaus Universitetas. Accessed 2019-10-24.
24. Gurulingappa, Harsha, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports." Journal of Biomedical Informatics, Elsevier, vol. 45, no. 5, pp. 885-892, October. Accessed 2019-10-28.
25. Healey, Antonette diPaolo. 2009. "Dictionary of Old English Corpus (DOE Corpus)." CoRD, Varieng, November 23. Updated 2011-03-17. Accessed 2019-10-25.
26. Iderhoff, Nicolas. 2019. "niderhoff/nlp-datasets." GitHub, October

18. Accessed 2019-10-25.

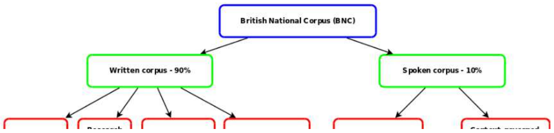
27. Johnston, Trevor. 2008. "From archive to corpus: transcription and annotation in the creation of signed language corpora." Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, pp. 16-29, November. Accessed 2019-10-25.
28. Kauhanen, Henri. 2011. "The Standard Corpus of Present-Day Edited American English (the Brown Corpus)." VARIENG, University of Helsinki, March 20. Accessed 2019-09-06.
29. LDC. 2019. "About LDC." Linguistic Data Consortium. Accessed 2019-10-24.
30. Marcus, Mitch. 2011. "A Brief History of the Penn Treebank." Center for Language and Speech Processing, Johns Hopkins University, February 15. Accessed 2019-09-06.
31. Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. "Treebank-3." LDC99T42, Philadelphia: Linguistic Data Consortium. Accessed 2019-10-24.
32. Mayo, Matthew. 2017. "Building a Wikipedia Text Corpus for Natural Language Processing." KDnuggets, November. Accessed 2019-10-23.
33. NLTK. 2019. "Corpus Readers." NLTK. Accessed 2019-10-23.
34. Nagel, Sebastian. 2019. "September 2019 crawl archive now available." Common Crawl, September 28. Accessed 2019-10-25.
35. Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme. 2012. "Annotated English Gigaword." LDC2012T21, Philadelphia: Linguistic Data Consortium, November 15. Accessed 2019-10-24.
36. Nivre, Joakim. 2007. "Treebanks." Chapter 11 in: Corpus Linguistics, Handbooks of Linguistics and Communication Science (HSK), De Gruyter. Accessed 2019-10-28.
37. O'Donnell, Matthew Brook. 2008. "Corpus Mark-up." UoL Summer Institute in Corpus Linguistics, July 01. Accessed 2019-10-23.
38. ODSC. 2019. "20 Open Datasets for Natural Language Processing." Medium, July 31. Accessed 2019-10-25.
39. OPUS. 2019. "Wikipedia." The Open Parallel Corpus. Accessed 2019-10-25.
40. Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. "English Gigaword Fifth Edition." LDC2011T07, Philadelphia: Linguistic Data Consortium, June 17. Accessed 2019-10-24.
41. Registry of Open Data. 2019. "Amazon Customer Reviews Dataset." Registry of Open Data on AWS. Accessed 2019-10-24.
42. Rennie, Jason. 2008. "20 Newsgroup." January 14. Accessed 2019-10-24.
43. Ruder, Sebastian. 2018a. "NLP's ImageNet moment has arrived." July 12. Accessed 2019-10-23.
44. Ruder, Sebastian. 2018b. "Semantic role labeling." NLP-progress, on GitHub, October 25. Accessed 2019-10-24.
45. Ruder, Sebastian. 2019a. "Constituency parsing." NLP-progress, on GitHub, September 14. Accessed 2019-10-24.
46. Ruder, Sebastian. 2019b. "Machine translation." NLP-progress, on GitHub, February 25. Accessed 2019-10-24.
47. Ruder, Sebastian. 2019c. "Part-of-speech tagging." NLP-progress, on GitHub, September 24. Accessed 2019-10-24.

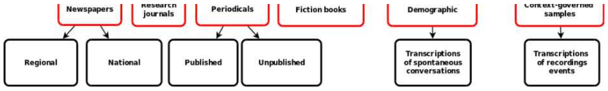
48. Ruder, Sebastian. 2019d. "Named entity recognition." NLP-progress, on GitHub, April 30. Accessed 2019-10-24.
49. Ruder, Sebastian. 2019e. "Language modeling." NLP-progress, on GitHub, October 06. Accessed 2019-10-24.
50. SQuAD. 2019. "SQuAD 2.0: The Stanford Question Answering Dataset." Stanford NLP Group. Accessed 2019-10-23.
51. SQuAD. 2019b. "Black_Death." SQuAD 2.0, Stanford NLP Group. Accessed 2019-10-23.
52. Silveira, Natalia, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. "A Gold Standard Dependency Corpus for English." Proceedings of the Ninth International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), pp. 2897-2904, May. Accessed 2019-10-25.
53. Sketch Engine. 2017. "Open Cambridge Learner Corpus (Uncoded)." Sketch Engine, March 15. Updated 2019-06-18. Accessed 2019-10-25.
54. Stanford NLP. 2019a. "Coreference Resolution." The Stanford Natural Language Processing Group. Accessed 2019-10-23.
55. UCREL. 2019. "UCREL Semantic Analysis System (USAS)." UCREL, University of Lancaster. Accessed 2019-10-28.
56. Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. "OntoNotes Release 5.0." LDC2013T19, Philadelphia: Linguistic Data Consortium, October 13. Accessed 2019-10-24.
57. Wikipedia. 2019. "List of text corpora." Wikipedia, July 26. Accessed 2019-10-24.
58. Wikipedia. 2019b. "British National Corpus." Wikipedia, October 22. Accessed 2019-10-24.
59. Word Frequency. 2019. "Word frequency data." Accessed 2019-10-23.
60. Yelp. 2019. "Yelp Open Dataset." Accessed 2019-10-24.
61. Zhang, Ye and Byron C. Wallace. 2017. "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification." Proceedings of the 8th International Joint Conference on Natural Language Processing, pp. 253–263, Taipei, Taiwan, November 27 – December 1. Accessed 2019-10-23.

Milestones

W. Nelson Francis and Henry Kučera at the Department of Linguistics, Brown University, publish a computer-readable general corpus to aid linguistic research on modern English. The corpus has 1 million words (500 samples of about 2000 words each). Revised editions appear later in 1971 and 1979. Called **Brown Corpus**, it inspires many other text corpora. The corpus with annotations is included in Treebank-3 (1999).

Linguistic Data Consortium (LDC) is formed to serve as a repository for NLP resources, including corpora. It's hosted at the University of Pennsylvania.





A 100-million corpus of British English called **BNC (British National Corpus)** is assembled between 1991 and 1994. It's balanced across genres. A follow-up task called **BNC2014** is started in 2014, which can help in understanding how language evolves. Spoken BNC2014 is released in September 2017. Written BNC2014 is expected to come out in 2019.

Penn Treebank-3 is released. It's based upon the original Treebank (1992) and its revised Treebank II (1995). This work started in 1989 at the University of Pennsylvania. Treebank-3 includes tagged/parsed Brown Corpus, 1 million words of 1989 WSJ material annotated in Treebank II style, tagged sample of ATIS-3, and tagged/parsed Switchboard Corpus. Apart from POS tags, the corpus includes chunk tags, relation tags and anchor tags. The **BLLIP 1987-89 WSJ Corpus Release 1** has 30 million words and supplements the WSJ section of Treebank-3.

Collected for the years 1990-2007, the **Corpus of Contemporary American English (COCA)** is released with 365 million words. By December 2017, it has 560 million words, adding 20 million each year. There's good balance of spoken, fiction, popular magazines, newspapers, and academic texts. It's been noted that COCA contains many common words that are missing in the American National Corpus (ANC), a corpus of 22 million words.

English Gigaword Fifth Edition is released by LDC. It comes from seven English newswire services. It has 4 billion words and takes up 26 gigabytes uncompressed. The first edition appeared in 2003. In November 2012, researchers at the John Hopkins University add syntactic and discourse structure annotations to this corpus after parsing more than 183 million sentences.

From digitized books, Google releases version 2 of **Google Books Ngrams**. Version 1 came out in July 2009. Only n-grams that appear more than 40 times are included. The corpus includes 1-gram to 5-grams. It includes many non-English languages as well. To experiment on small sets of phrases, researchers can try out the online Google Books Ngram Viewer.

As a corpus for informal genre, **English Web Treebank (EWT)** is released by LDC. This includes content from weblogs, reviews, question-answers, newsgroups, and email. It has about 250K word-level tokens and 16K sentence-level tokens. It's annotated for POS and syntactic structure. This includes Enron Corporation emails from 1999-2002. In 2014, Silveira et al. provide annotation of syntactic dependencies for this corpus that can be used to train dependency parsers.

Common Crawl publishes 240 TiB of uncompressed data from 2.55 billion web pages. Of these, 1 billion URLs were not present in previous crawls. Common Crawl started in 2008. In 2013, they moved from ARC to Web ARChive (WARC) file format. WAT files contain the metadata. WET file contain plaintext of the WARC files.

Tags

See Also

- Speech Corpus for NLP
- Neural Networks for NLP
- Word Embedding

- Open Data
- Structured vs Unstructured Data
- Data Mining

Further Reading

1. Iderhoff, Nicolas. 2019. "niderhoff/nlp-datasets." GitHub, October 18. Accessed 2019-10-25.
2. Ali, Meiryum. 2019. "The Best 25 Datasets for Natural Language Processing." Lionbridge, July 09. Accessed 2019-10-23.
3. Brownlee, Jason. 2017. "Datasets for Natural Language Processing." Machine Learning Mastery, September 27. Updated 2019-08-07. Accessed 2019-10-23.
4. Gries, Stefan Th. and Andrea L. Berez. 2017. "Linguistic Annotation in/for Corpus Linguistics." Chapter in: N. Ide and J. Pustejovsky (eds.), Handbook of Linguistic Annotation, Springer Science+Business Media Dordrecht, pp. 379-409. Accessed 2019-10-23.

Article Stats

Author-wise Stats for Article Edits

Author

No. of Edits

No. of Chats

DevCoins

Cite As

Devopedia. 2019. "Text Corpus for NLP." Version 3, October 28. Accessed 2019-12-10. https://devopedia.org/text-corpus-for-nlp