Natural Language Processing Introduction

Unit 2: Corpora Analysis

February 2020

Text Corpus

Definition

A text corpus is a very large collection of text produced by **real users of the language** and used to analyse how words, phrases and language in general are used. A corpus is also be used for generating various language databases used in software development such as predictive keyboards, spell check, grammar correction, text/speech understanding systems, text-to-speech modules and many others.

Why are they relevant?

- Tools to preserve human knowledge.

Why are they relevant?

- Tools to preserve human knowledge.
- Key elements of NLP systems.

Why are they relevant?

- Tools to preserve human knowledge.
- Key elements of NLP systems.
- Useful for cross lingual learning.

Why are they relevant?

- Tools to preserve human knowledge.
- Key elements of NLP systems.
- Useful for cross lingual learning.
- Experimental spaces of unstructured data.

Monolingual corpus

Monolingual corpus is the most frequent type of corpus. It contains texts in one language only. The corpus is usually tagged for parts of speech and is used by a wide range of users for various tasks from highly practical ones, e.g. checking the correct usage of a word or looking up the most natural word combinations, to scientific use, e.g. identifying frequent patterns or new trends in language.

Parallel corpus

A parallel corpus consists of two monolingual corpora. One corpus is the translation of the other. For example, a novel and its translation or a translation memory of a CAT tool could be used to build a parallel corpus. Both languages need to be aligned, i.e. corresponding segments, usually sentences or paragraphs, need to be matched. The user can then search for all examples of a word or phrase in one language and the results will be displayed together with the corresponding sentences in the other language. The user can then observe how the search word or phrase is translated.

Multilingual corpus

A multilingual corpus is very similar to a parallel corpus. The two terms are often used interchangeably. A multilingual corpus contains texts in several languages which are all translations of the same text and are aligned in the same way as parallel corpora.
When only two languages are selected, a multilingual corpus behaves as a parallel corpus. The user can also decide to work with one language to use it as a monolingual corpus.
The terms parallel and multilingual are sometimes used interchangeably.

Comparable corpus

A comparable corpus is a set of two or more monolingual corpora whose texts relate to the same topic., however, they are not translations of each other, and therefore, there are not aligned. When users search these corpora they can use the fact, that the corpora also have the same metadata.

Learner corpus

A learner corpus is a corpus of texts produced by learners of a language. The corpus is used to study the mistakes and problems learners have when learning a foreign language.

Diachronic corpus

A diachronic corpus is a corpus containing texts from different periods and is used to study the development or change in language. This corpus can be used to indentify trends, which identifies words whose usage changes the most of the selected period of time.

Specialized corpus

A specialized corpus contains texts limited to one or more subject areas, domains, topics etc. Such corpus is used to study how the specialized language is used.

Multimedia corpus

A multimedia corpus contains texts which are enhanced with audio or visual materials or other type of multimedia content. It can include images, text, audio and video simultaneously.

Where to get the data?

When we are talking about machine learning tasks, one of the most crucial parts to build any systems is the data. NLP data comes in a variety of forms, depending on the task we want to perform. But the main question is:

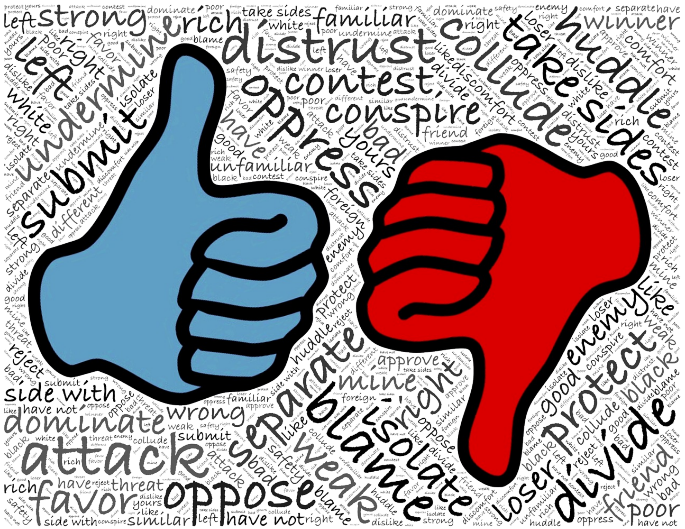
Where to get the data?

When we are talking about machine learning tasks, one of the most crucial parts to build any systems is the data. NLP data comes in a variety of forms, depending on the task we want to perform. But the main question is:

Where do I get the data?

NLP Benchmarks

NLP benchmarks are a set of datasets and tasks which are ment to compare the performance of different NLP general models. They result useful for scientist and practitioners. Benchmarks often offers a set of datasets and tasks to evelaute general NLP models.



Glue

General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing NLU systems comprising:

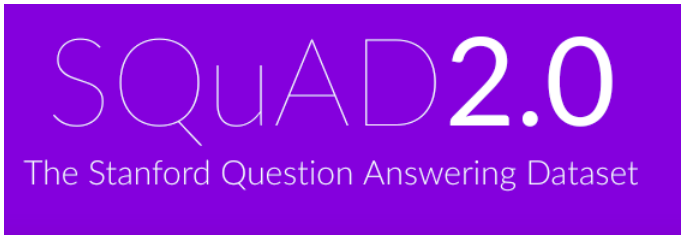
- A benchmark of nine sentence or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty.
- A diagnostic dataset designed to evaluate and analyze model performance.
- A public leaderboard for tracking performance.

SuperGLUE

GLUE benhmark enhanced. It took into account the lessons learnt from original GLUE benchmark and presented SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard.

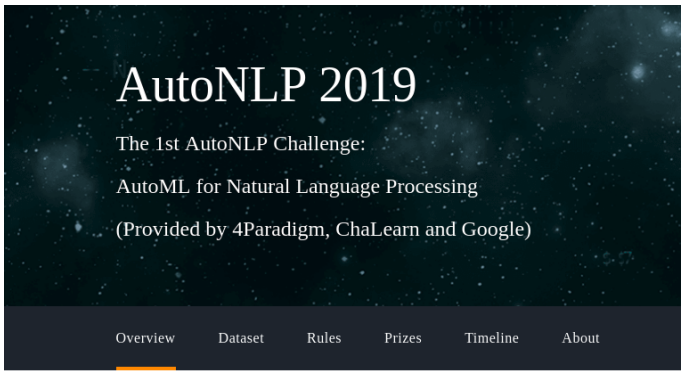
SQuAD

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.



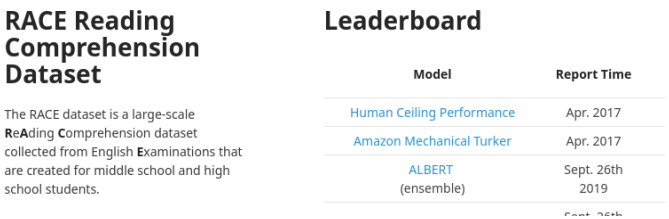
AutoNLP

In this challenge, participants are invited to develop AutoNLP solutions to multi-class text categorization problems. The provided datasets are in the form of text. Five practice datasets, which can be downloaded by the participants, are provided to the participants so that they can develop their AutoNLP solutions offline. Besides that, another five validation datasets are also provided to participants to evaluate the public leaderboard scores of their AutoNLP solutions.



Race

The RACE dataset is a large-scale ReAding Comprehension dataset collected from English Examinations that are created for middle school and high school students.



RACE Reading Comprehension Dataset

The RACE dataset is a large-scale **Re**A**d**ing **C**omprehension dataset collected from English **E**xaminations that are created for middle school and high school students.

Leaderboard

Model	Report Time
Human Ceiling Performance	Apr. 2017
Amazon Mechanical Turker	Apr. 2017
ALBERT (ensemble)	Sept. 26th 2019

Glue corpora example:

- The Corpus of Linguistic Acceptability
- The Stanford Sentiment Treebank
- Microsoft Research Paraphrase Corpus
- Semantic Textual Similarity Benchmark
- Quora Question Pairs
- MultiNLI Matched
- MultiNLI Mismatched
- Question NLI
- Recognizing Textual Entailment
- Winograd NLI
- Diagnostics Main

Benchmarks corpora

Links to interesting NLP benchmarks:

- Glue: https://gluebenchmark.com/leaderboard/
- Superglue: https://super.gluebenchmark.com/leaderboard
- Race: http://www.qizhexie.com/data/RACE_leaderboard.html
- AutoNLP https://www.4paradigm.com/competition/autoNLP2019
- SQuAD https://rajpurkar.github.io/SQuAD-explorer/

Kaggle

Kaggle is a collaborative platform for Data Scientist, where interesting data and data science competitions are hosted and evaluated. Important companies at global level for competitions for Data scientist around the world, and is a great tool for data science community to learn, share and get valuable datasets.

Kaggle

Zeeshan-ul-hassan Usmani

What is Kaggle, Why I Participate, What is the Impact?

posted in Getting Started 2 years ago

15

Kaggle is an AirBnB for Data Scientists – this is where they spend their nights and weekends. It's a crowd-sourced platform to attract, nurture, train and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over 536,000 active members from 194 countries and it receives close to 150,000 submissions per month. Started from Melbourne, Australia Kaggle moved to Silicon Valley in 2011, raised some 11 million dollars from the likes of Hal Varian (Chief Economist at Google), Max Levchin (Paypal), Index and Khosla Ventures and then ultimately been acquired by the Google in March of 2017. Kaggle is the number one stop for data science enthusiasts all around the world who compete for prizes and boost their Kaggle rankings. There are only 94 Kaggle Grandmasters in the world to this date.

Corpus in spanish

Corpus in spanish

Corpus in spanish

Best corpus in spanish

- El corpus del español: A 100 million word corpus from over 20,000 Spanish texts spanning from 1200 to the 1900s.
- MAS Corpus (Marketing Analysis in Spanish): Contains manually tagged Twitter posts in Spanish for marketing.
- 120 Million Word Spanish Corpus: A medium sized corpus containing 120 million words of modern Spanish taken from the Spanish Language Wikipedia in 2010.
- The TV News Archive: Over 705,000 captioned and searchable news programs from over 4 years of U.S. television networks.
- Spanish norms for photographs: 140 color images normed by over one hundred native Spanish speakers.

Retrieved from https://lionbridge.ai/datasets/22-best-spanish-language-datasets-for-machine-learning/

Corpus in spanish

Corpus analysis

Once I have obtained a corpus, which are the next stepsr to have? How should I start the analysis?

- Check the files: Which is the format of the files? How much they weight? Which codification does they have?
- Check the structure: How the information is distributed? How the topics are gathered? Which annotations it has? Are there any classification hints in the data?
- Check the data: Checkout random examples of the corpora. Which treatment they need? What kind of preprocessing you will have to make with them?

Assignments

Assignment 5: Download and describe a text Dataset:

- Download the SQuAD 2.0 Dataset
- Report the files in the Dataset
- Report the structure of the Dataset
- Extract samples of tha dataset and describe what kind of pre processing you need to perform in the given Dataset

Let's code

References

[1] Jurafsky D., Martin J.: Speech and Language Processing 2nd. ed. (2009).

[2] Bird S., Klein E., Loper, E.: Natural Language Processing with Python, (2009). ISBN: 978-0-596-51649-9

[3] Indurkhya N., Damerau F. Handbook of Natural Language Processing, Second Edition (2010). ISBN: 978-1-4200-8593-8

[4] Kao, A., Poteet S. Natural Language Processing and Text Mining (2007). ISBN: 78-1-84628-175-4

[5] Sipser, M. Introduction to the theory of computation (2013). ISBN: 978-1-133-18779-0

[6] https://www.sketchengine.eu/corpora-and-languages/corpus-types/