BIGBIO: A Framework for Data-Centric Biomedical Natural Language Processing

Preprint. Under review.

A Example Data Card

We generated data cards for all BIGBIO datasets. We include an example dataset from each schema type to illustrate data cards for different tasks. A PDF of all content is available on our project homepage.

AnEM Data Card

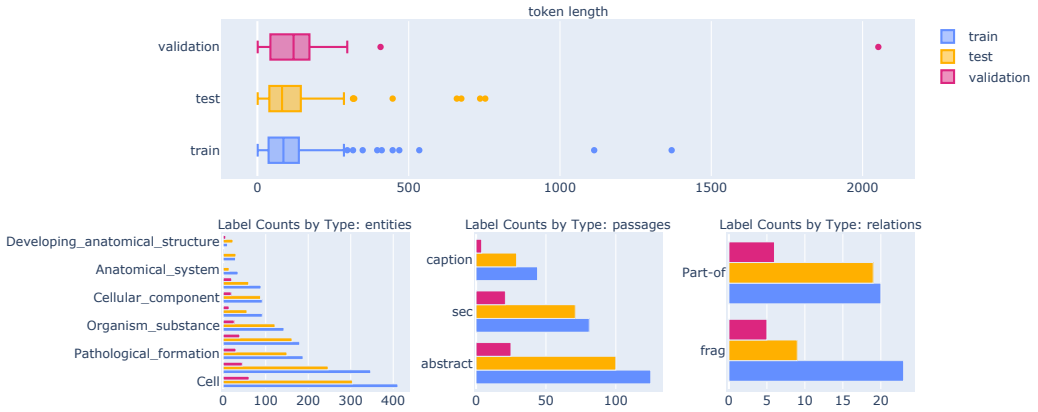

Figure 1: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: AnEM corpus is a domain- and species-independent resource manually annotated for anatomicalentity mentions using a fine-grained classification system. The corpus consists of 500 documents(over 90,000 words) selected randomly from citation abstracts and full-text papers withthe aim of making the corpus representative of the entire available biomedical scientificliterature. The corpus annotation covers mentions of both healthy and pathological anatomicalentities and contains over 3,000 annotated mentions.

Homepage: `http://www.nactem.ac.uk/anatomy/`

URL: `http://www.nactem.ac.uk/anatomy/`

Licensing: Creative Commons Attribution Share Alike 3.0 Unported

Languages: English

Tasks: named entity recognition, relation extraction, coreference resolution

Schemas: `KB`

Splits: train, test, validation

AnatEM Data Card

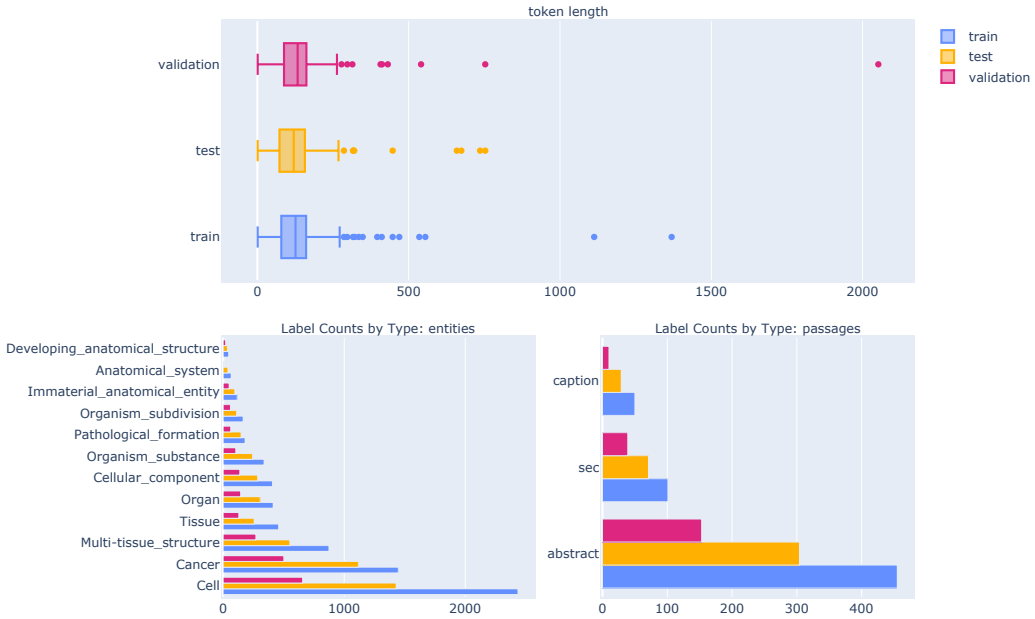

Figure 2: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The extended Anatomical Entity Mention corpus (AnatEM) consists of 1212 documents (approx. 250,000 words)manually annotated to identify over 13,000 mentions of anatomical entities. Each annotation is assigned oneof 12 granularity-based types such as Cellular component, Tissue and Organ, defined with reference to theCommon Anatomy Reference Ontology.

Homepage: `http://nactem.ac.uk/anatomytagger/#AnatEM`

URL: `http://nactem.ac.uk/anatomytagger/#AnatEM`

Licensing: Creative Commons Attribution Share Alike 3.0 Unported

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

AskAPatient Data Card

Figure 3: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The AskAPatient dataset contains medical concepts written on social media mapped to how they are formally written in medical ontologies (SNOMED-CT and AMT).

Homepage: `https://zenodo.org/record/55013`

URL: `https://zenodo.org/record/55013`

Licensing: Creative Commons Attribution 4.0 International

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train, validation, test

BC5CDR Data Card

Figure 4: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioCreative V Chemical Disease Relation (CDR) dataset is a large annotated text corpus ofhuman annotations of all chemicals, diseases and their interactions in 1,500 PubMed articles.

Homepage: `http://www.biocreative.org/tasks/biocreative-v/track-3-cdr/`

URL: `http://www.biocreative.org/tasks/biocreative-v/track-3-cdr/`

Licensing: Public Domain Mark 1.0

Languages: English

Tasks: named entity recognition, relation extraction, named entity disambiguation

Schemas: `KB`

Splits: train, test, validation

BC7-LitCovid Data Card

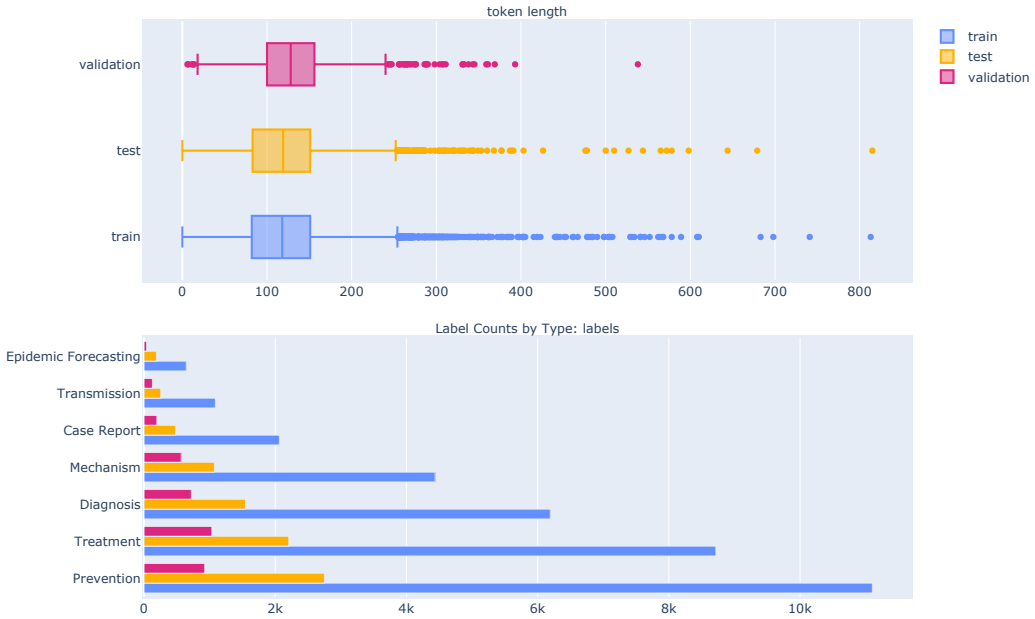

Figure 5: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The training and development datasets contain the publicly-availabletext of over 30 thousand COVID-19-related articles and their metadata(e.g., title, abstract, journal). Articles in both datasets have beenmanually reviewed and articles annotated by in-house models.

Homepage: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/`

URL: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/`

Licensing: License information unavailable

Languages: English

Tasks: text classification

Schemas: `TEXT`

Splits: train, test, validation

Bio-SimVerb Data Card

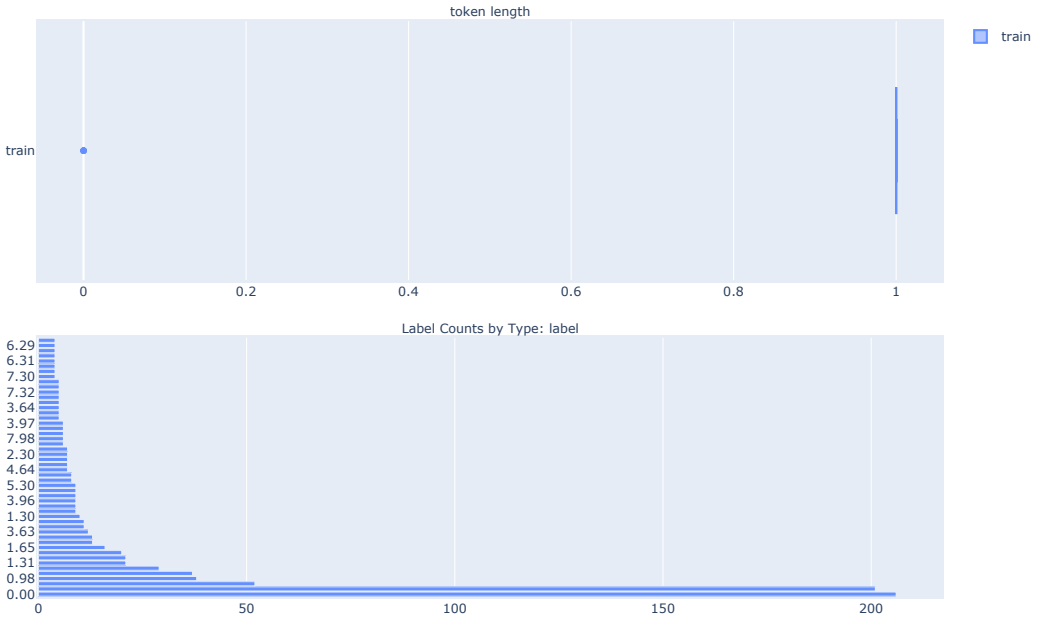

Figure 6: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: This repository contains the evaluation datasets for the paperBio-SimVerb and Bio-SimLex: Wide-coverage Evaluation Sets of WordSimilarity in Biomedicine by Billy Chiu, Sampo Pyysalo and Anna Korhonen.

Homepage: `https://github.com/cambridgeltl/bio-simverb` **URL:** `https://github.com/cambridgeltl/bio-simverb`

Licensing: License information unavailable

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

Bio-SimLex Data Card

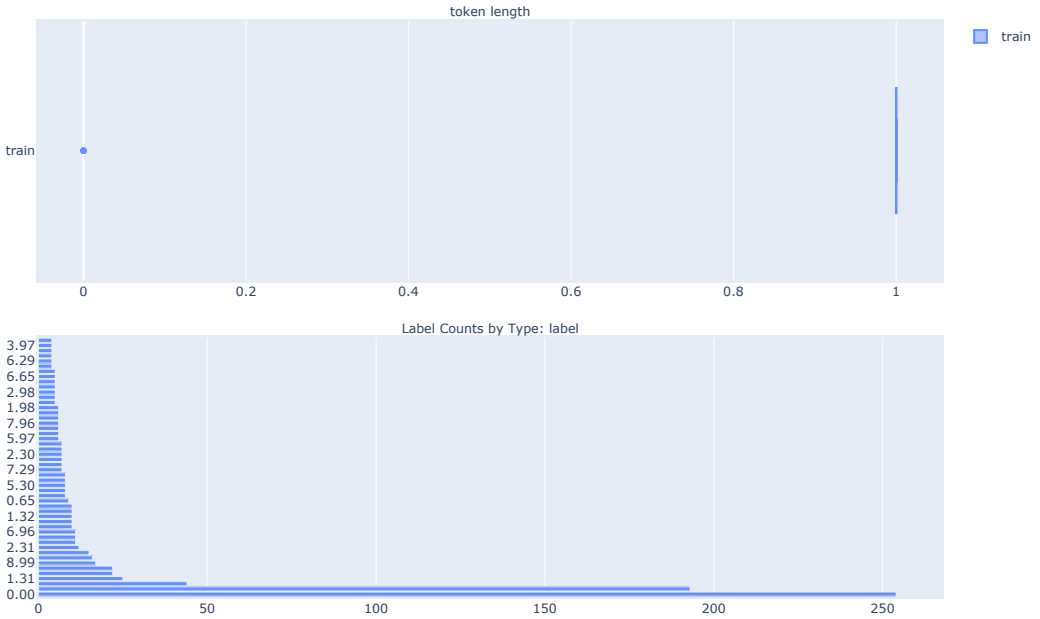

Figure 7: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Bio-SimLex enables intrinsic evaluation of word representations. This evaluation can serve as a predictor ofperformance on various downstream tasks in the biomedical domain. The results on Bio-SimLex using standardword representation models highlight the importance of developing dedicated evaluation resources for NLP in biomedicinefor particular word classes (e.g. verbs).

Homepage: `https://github.com/cambridgeltl/bio-simverb`

URL: `https://github.com/cambridgeltl/bio-simverb`

Licensing: License information unavailable

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

MESINESP 2021 Subtrack 1 (All) Data Card

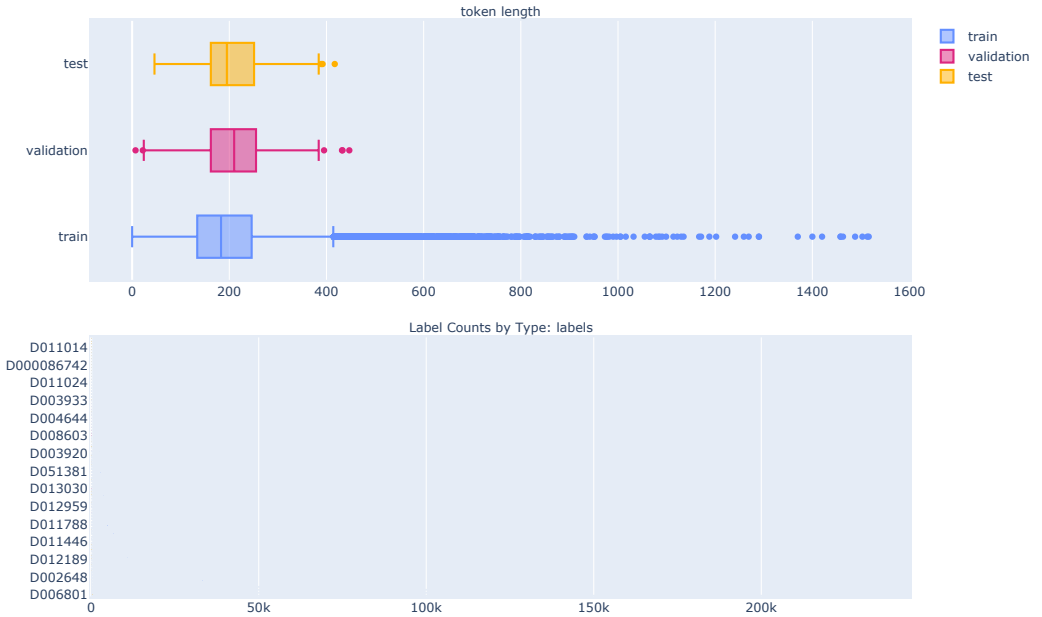

Figure 8: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The main aim of MESINESP2 is to promote the development of practically relevant semantic indexing tools for biomedical content in non-English language.We have generated a manually annotated corpus, where domain experts have labeled a set of scientific literature, clinical trials, and patent abstracts.All the documents were labeled with DeCS descriptors, which is a structured controlled vocabulary created by BIREME to index scientific publications on BvSalud, the largest database of scientific documents in Spanish, which hosts records from the databases LILACS, MEDLINE, IBECS, among others. MESINESP track at BioASQ9 explores the efficiency of systems for assigning DeCS to different types of biomedical documents. To that purpose, we have divided the task into three subtracks depending on the document type. Then, for each one we generated an annotated corpus which was provided to participating teams:

[Subtrack 1 corpus] MESINESP-L – Scientific Literature: It contains all Spanish records from LILACS and IBECS databases at the Virtual Health Library(VHL) with non-empty abstract written in Spanish.

[Subtrack 2 corpus] MESINESP-T- Clinical Trials contains records from Registro Español de Estudios Clínicos (REEC). REEC doesn't provide documents with the structure title/abstract needed in BioASQ, for that reason we have built artificial abstracts based on the content available in the data crawled using the REEC API.

[Subtrack 3 corpus] MESINESP-P – Patents: This corpus includes patents in Spanish extracted from Google Patents which have the IPC code "A61P" and "A61K31".In addition, we also provide a set of complementary data such as: the DeCS terminology file, a silver standard with the participants' predictions to the task background set and the entities of medications, diseases, symptoms and medical procedures extracted from the BSC NERs documents. **Homepage:** `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

URL: `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `TEXT`

Splits: train, validation, test

MESINESP 2021 Subtrack 1 (Only Articles) Data Card

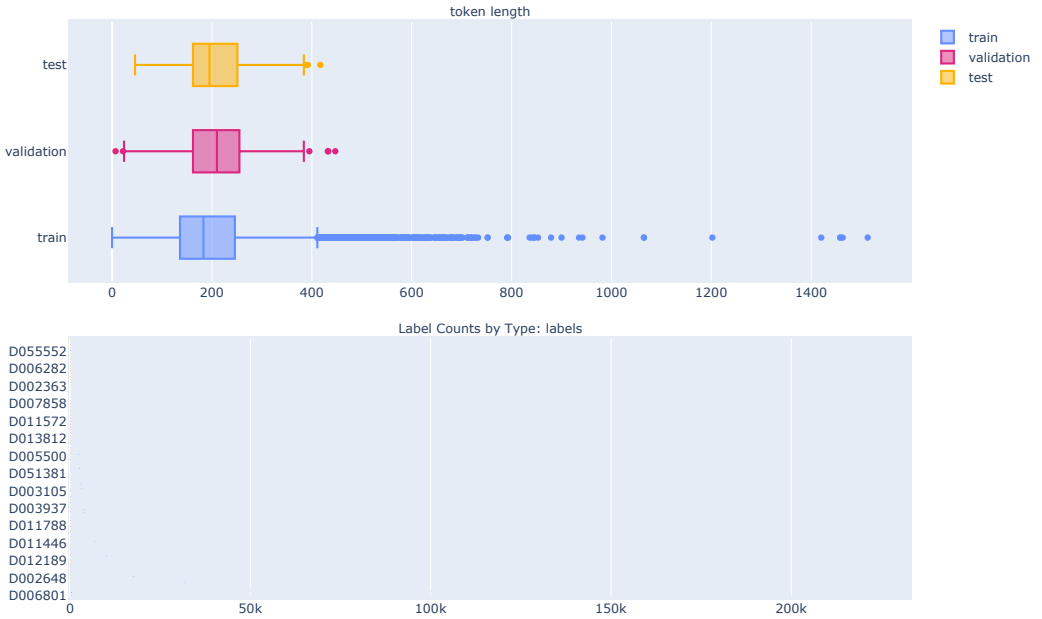

Figure 9: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The main aim of MESINESP2 is to promote the development of practically relevant semantic indexing tools for biomedical content in non-English language.We have generated a manually annotated corpus, where domain experts have labeled a set of scientific literature, clinical trials, and patent abstracts.All the documents were labeled with DeCS descriptors, which is a structured controlled vocabulary created by BIREME to index scientific publications on BvSalud, the largest database of scientific documents in Spanish, which hosts records from the databases LILACS, MEDLINE, IBECS, among others. MESINESP track at BioASQ9 explores the efficiency of systems for assigning DeCS to different types of biomedical documents. To that purpose, we have divided the task into three subtracks depending on the document type. Then, for each one we generated an annotated corpus which was provided to participating teams:

[Subtrack 1 corpus] MESINESP-L – Scientific Literature: It contains all Spanish records from LILACS and IBECS databases at the Virtual Health Library(VHL) with non-empty abstract written in Spanish.

[Subtrack 2 corpus] MESINESP-T- Clinical Trials contains records from Registro Español de Estudios Clínicos (REEC). REEC doesn't provide documents with the structure title/abstract needed in BioASQ, for that reason we have built artificial abstracts based on the content available in the data crawled using the REEC API.

[Subtrack 3 corpus] MESINESP-P – Patents: This corpus includes patents in Spanish extracted from Google Patents which have the IPC code "A61P" and "A61K31".In addition, we also provide a set of complementary data such as: the DeCS terminology file, a silver standard with the participants' predictions to the task background set and the entities of medications, diseases, symptoms and medical procedures extracted from the BSC NERs documents. **Homepage:** `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

URL: `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `TEXT`

Splits: train, validation, test

MESINESP 2021 Subtrack 2 Data Card

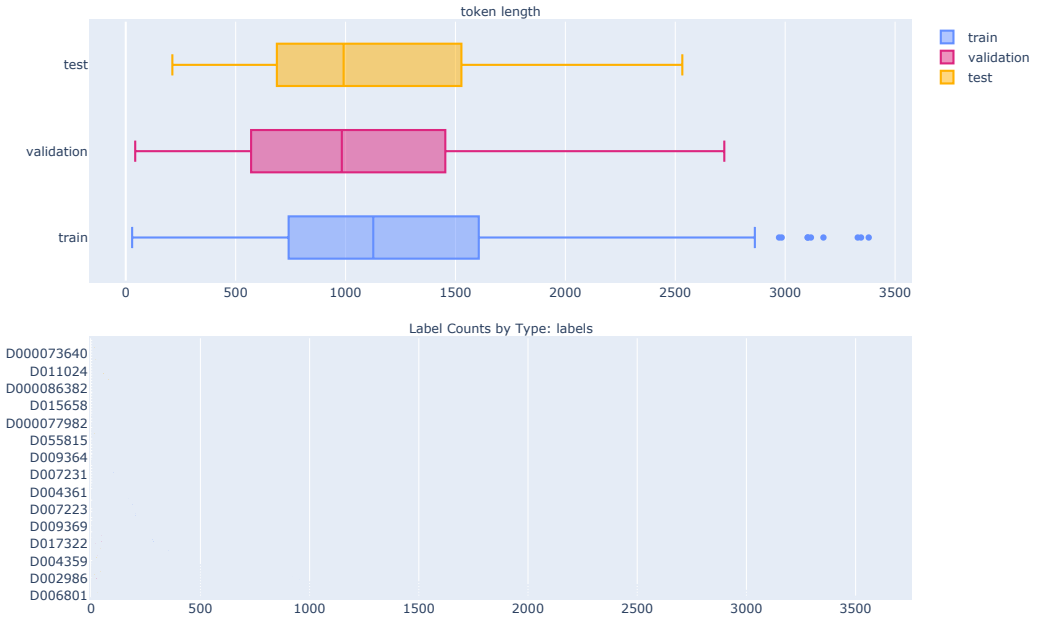

Figure 10: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The main aim of MESINESP2 is to promote the development of practically relevant semantic indexing tools for biomedical content in non-English language.We have generated a manually annotated corpus, where domain experts have labeled a set of scientific literature, clinical trials, and patent abstracts.All the documents were labeled with DeCS descriptors, which is a structured controlled vocabulary created by BIREME to index scientific publications on BvSalud, the largest database of scientific documents in Spanish, which hosts records from the databases LILACS, MEDLINE, IBECS, among others. MESINESP track at BioASQ9 explores the efficiency of systems for assigning DeCS to different types of biomedical documents. To that purpose, we have divided the task into three subtracks depending on the document type. Then, for each one we generated an annotated corpus which was provided to participating teams:

[Subtrack 1 corpus] MESINESP-L – Scientific Literature: It contains all Spanish records from LILACS and IBECS databases at the Virtual Health Library(VHL) with non-empty abstract written in Spanish.

[Subtrack 2 corpus] MESINESP-T- Clinical Trials contains records from Registro Español de Estudios Clínicos (REEC). REEC doesn't provide documents with the structure title/abstract needed in BioASQ, for that reason we have built artificial abstracts based on the content available in the data crawled using the REEC API.

[Subtrack 3 corpus] MESINESP-P – Patents: This corpus includes patents in Spanish extracted from Google Patents which have the IPC code "A61P" and "A61K31".In addition, we also provide a set of complementary data such as: the DeCS terminology file, a silver standard with the participants' predictions to the task background set and the entities of medications, diseases, symptoms and medical procedures extracted from the BSC NERs documents. **Homepage:** `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

URL: `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `TEXT`

Splits: train, validation, test

MESINESP 2021 Subtrack 3 Data Card

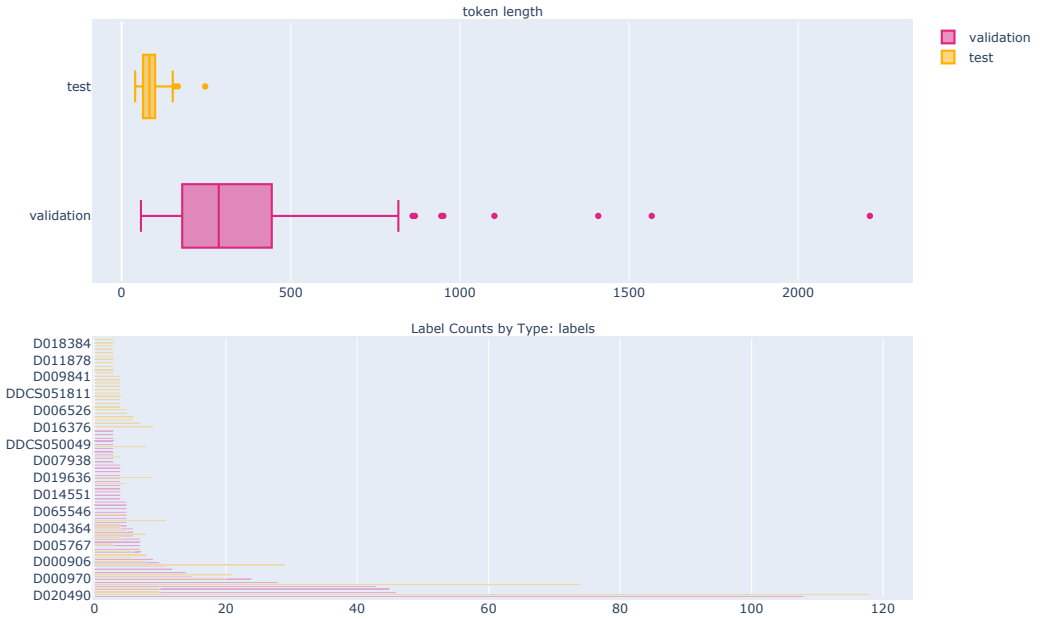

Figure 11: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The main aim of MESINESP2 is to promote the development of practically relevant semantic indexing tools for biomedical content in non-English language.We have generated a manually annotated corpus, where domain experts have labeled a set of scientific literature, clinical trials, and patent abstracts.All the documents were labeled with DeCS descriptors, which is a structured controlled vocabulary created by BIREME to index scientific publications on BvSalud, the largest database of scientific documents in Spanish, which hosts records from the databases LILACS, MEDLINE, IBECS, among others. MESINESP track at BioASQ9 explores the efficiency of systems for assigning DeCS to different types of biomedical documents. To that purpose, we have divided the task into three subtracks depending on the document type. Then, for each one we generated an annotated corpus which was provided to participating teams:

[Subtrack 1 corpus] MESINESP-L – Scientific Literature: It contains all Spanish records from LILACS and IBECS databases at the Virtual Health Library(VHL) with non-empty abstract written in Spanish.

[Subtrack 2 corpus] MESINESP-T- Clinical Trials contains records from Registro Español de Estudios Clínicos (REEC). REEC doesn't provide documents with the structure title/abstract needed in BioASQ, for that reason we have built artificial abstracts based on the content available in the data crawled using the REEC API.

[Subtrack 3 corpus] MESINESP-P – Patents: This corpus includes patents in Spanish extracted from Google Patents which have the IPC code "A61P" and "A61K31".In addition, we also provide a set of complementary data such as: the DeCS terminology file, a silver standard with the participants' predictions to the task background set and the entities of medications, diseases, symptoms and medical procedures extracted from the BSC NERs documents. **Homepage:** `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

URL: `https://zenodo.org/record/5602914#.YhSXJ5PMKWt`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `TEXT`

Splits: train, validation, test

BioInfer Data Card

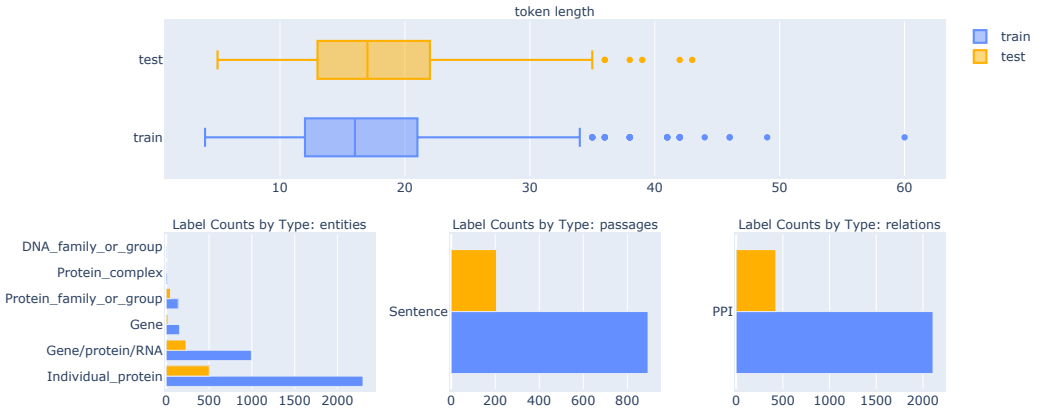

Figure 12: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: A corpus targeted at protein, gene, and RNA relationships which serves as a resource for the development of information extraction systems and their components such as parsers and domain analyzers. Currently, the corpus contains 1100 sentences from abstracts of biomedical research articles annotated for relationships, named entities, as well as syntactic dependencies.

Homepage: `https://github.com/metalrt/ppi-dataset`

URL: `https://github.com/metalrt/ppi-dataset`

Licensing: Creative Commons Attribution 2.0 Generic

Languages: English

Tasks: relation extraction, named entity recognition

Schemas: `KB`

Splits: train, test

BiologyHowWhy corpus Data Card

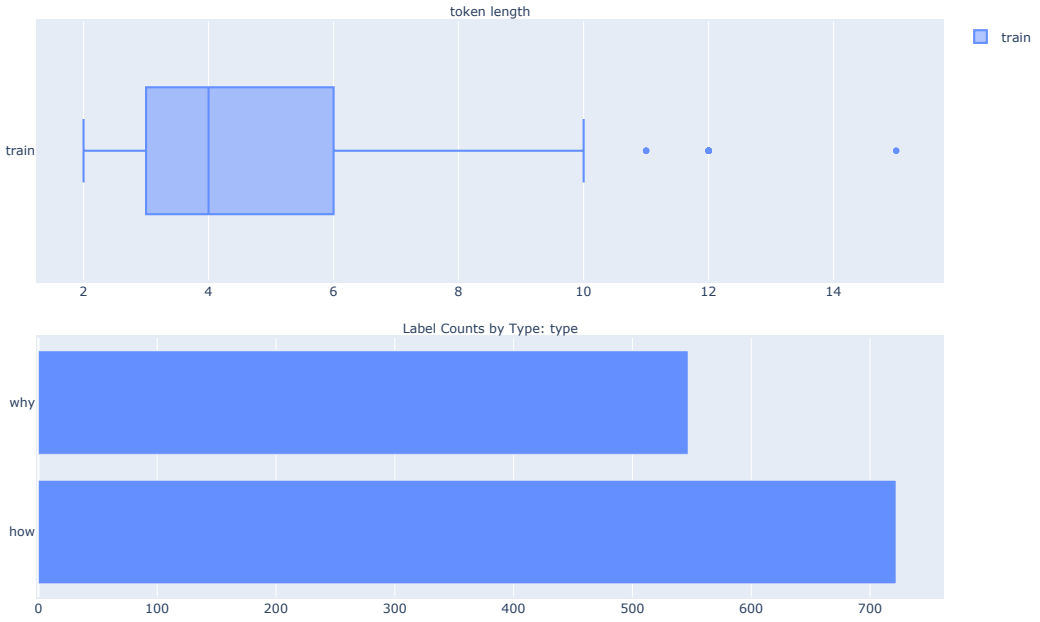

Figure 13: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: This dataset consists of 185 "how" and 193 "why" biology questions authored by a domain expert, with one or more gold answer passages identified in an undergraduate textbook. The expert was not constrained in any way during the annotation process, so gold answers might be smaller than a paragraph or span multiple paragraphs. This dataset was used for the question-answering system described in the paper "Discourse Complements Lexical Semantics for Non-factoid Answer Reranking" (ACL 2014). **Homepage:** `https://allenai.org/data/biology-how-why-corpus`

URL: `https://allenai.org/data/biology-how-why-corpus`

Licensing: License information unavailable

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train

BIOMRC Large A Data Card

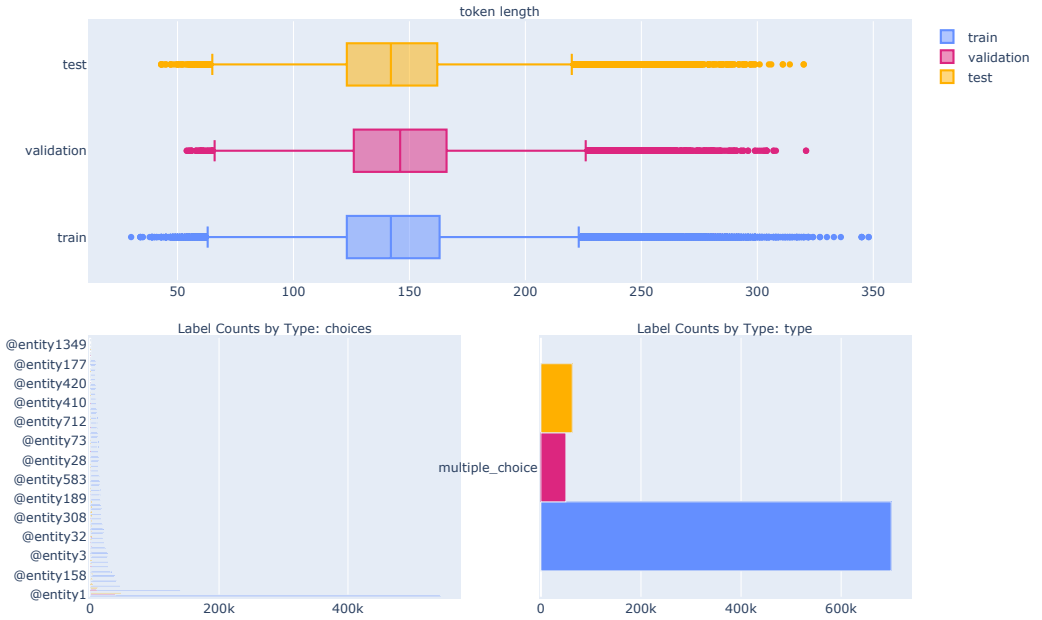

Figure 14: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We introduce BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments show that simple heuristics do not perform well on the new dataset and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Non-expert human performance is also higher on the new dataset compared to BIOREAD, and biomedical experts perform even better. We also introduce a new BERT-based MRC model, the best version of which substantially outperforms all other methods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make the new dataset available in three different sizes, also releasing our code, and providing a leader board.

Homepage: `https://github.com/PetrosStav/BioMRC_code`

URL: `https://github.com/PetrosStav/BioMRC_code`

Licensing: License information unavailable

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, validation, test

BIOMRC Large B Data Card

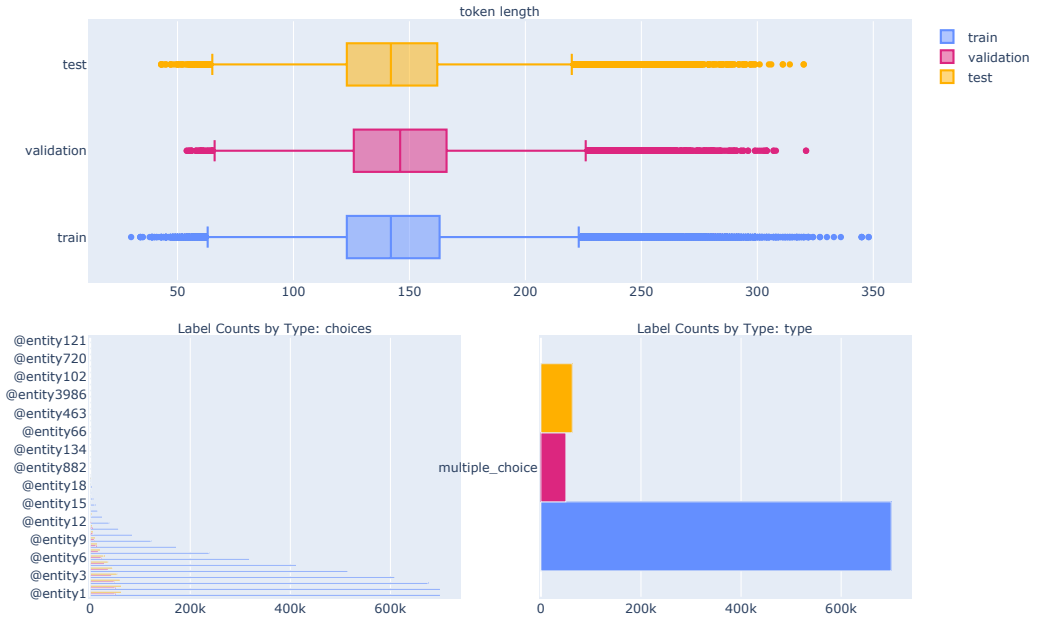

Figure 15: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We introduce BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments show that simple heuristics do not perform well on the new dataset and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Non-expert human performance is also higher on the new dataset compared to BIOREAD, and biomedical experts perform even better. We also introduce a new BERT-based MRC model, the best version of which substantially outperforms all other methods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make the new dataset available in three different sizes, also releasing our code, and providing a leader board.

Homepage: `https://github.com/PetrosStav/BioMRC_code`

URL: `https://github.com/PetrosStav/BioMRC_code`

Licensing: License information unavailable

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, validation, test

BIOMRC Small A Data Card

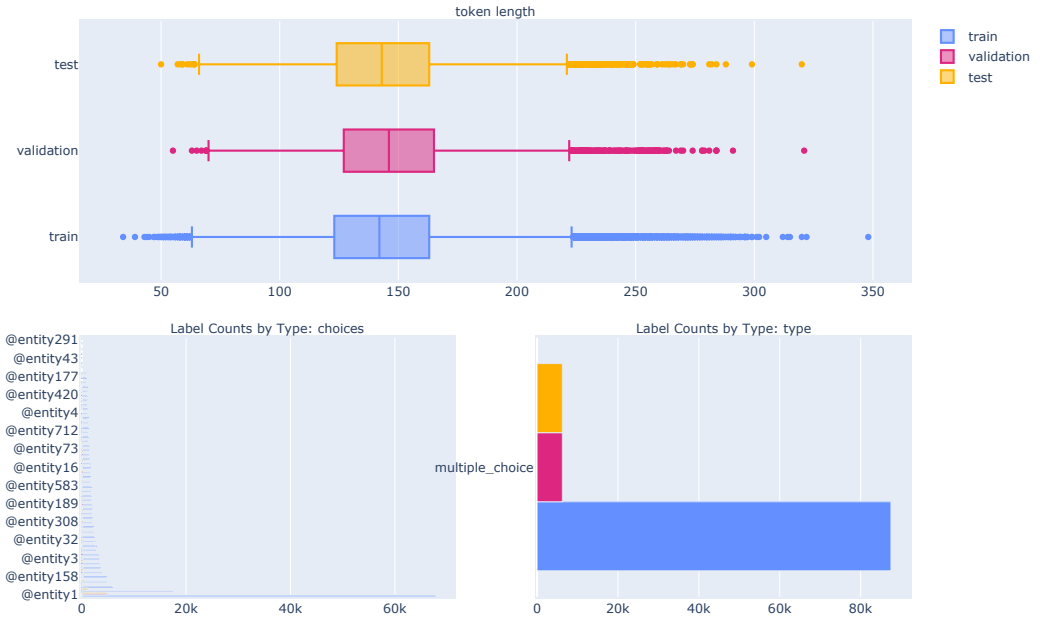

Figure 16: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We introduce BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments show that simple heuristics do not perform well on the new dataset and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Non-expert human performance is also higher on the new dataset compared to BIOREAD, and biomedical experts perform even better. We also introduce a new BERT-based MRC model, the best version of which substantially outperforms all other methods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make the new dataset available in three different sizes, also releasing our code, and providing a leader board.

Homepage: `https://github.com/PetrosStav/BioMRC_code`

URL: `https://github.com/PetrosStav/BioMRC_code`

Licensing: License information unavailable

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, validation, test

BIOMRC Small B Data Card

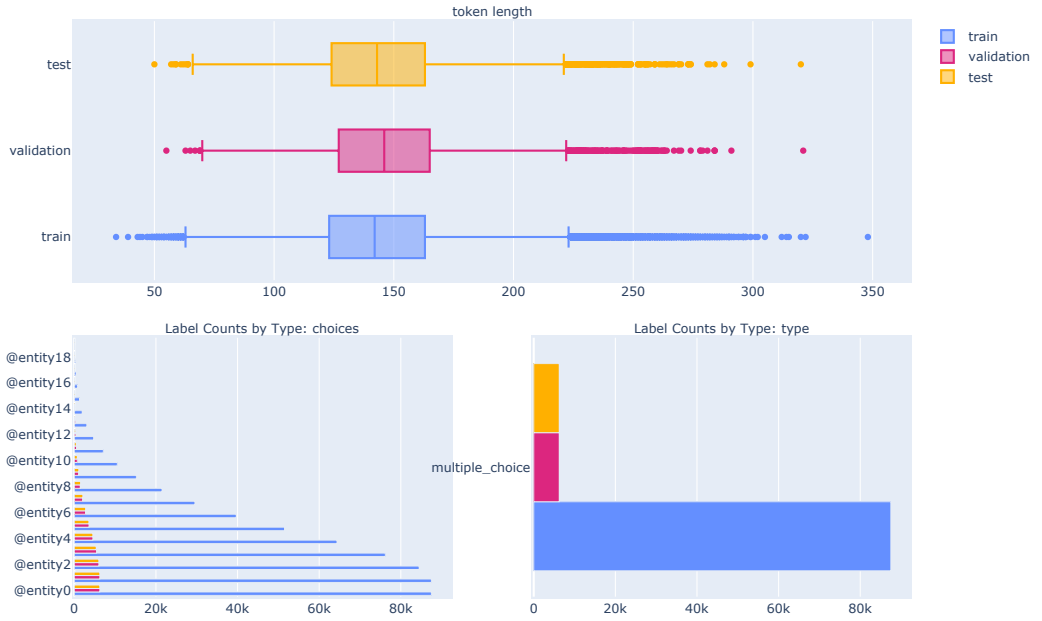

Figure 17: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We introduce BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments show that simple heuristics do not perform well on the new dataset and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Non-expert human performance is also higher on the new dataset compared to BIOREAD, and biomedical experts perform even better. We also introduce a new BERT-based MRC model, the best version of which substantially outperforms all other methods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make the new dataset available in three different sizes, also releasing our code, and providing a leader board.

Homepage: `https://github.com/PetrosStav/BioMRC_code`

URL: `https://github.com/PetrosStav/BioMRC_code`

Licensing: License information unavailable

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, validation, test

BIOMRC Tiny A Data Card

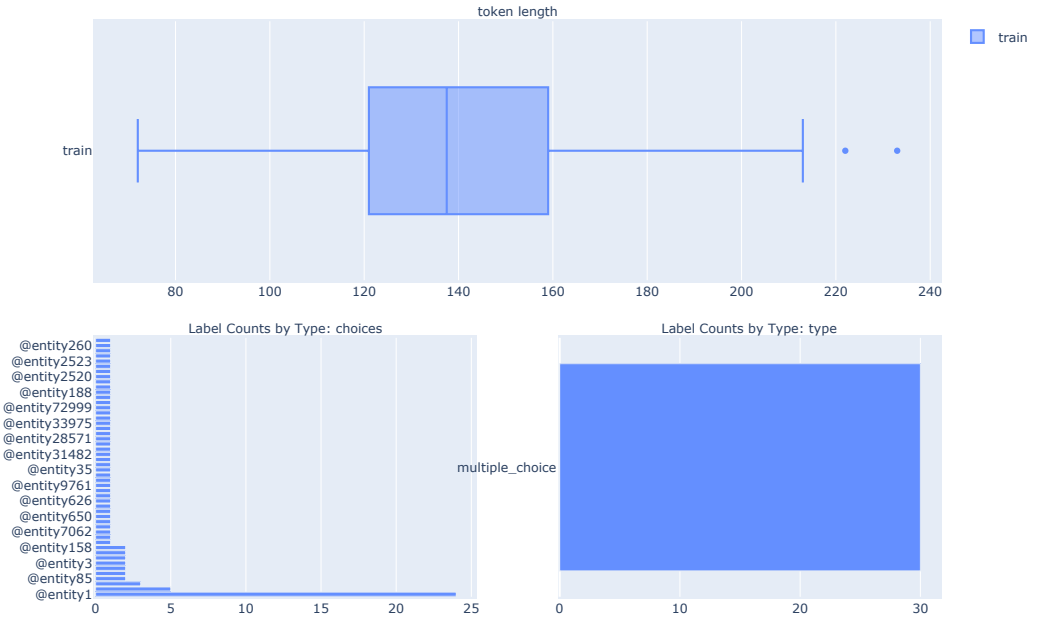

Figure 18: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We introduce BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments show that simple heuristics do not perform well on the new dataset and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Non-expert human performance is also higher on the new dataset compared to BIOREAD, and biomedical experts perform even better. We also introduce a new BERT-based MRC model, the best version of which substantially outperforms all other methods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make the new dataset available in three different sizes, also releasing our code, and providing a leader board.

Homepage: `https://github.com/PetrosStav/BioMRC_code`

URL: `https://github.com/PetrosStav/BioMRC_code`

Licensing: License information unavailable

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train

BIOMRC Tiny B Data Card

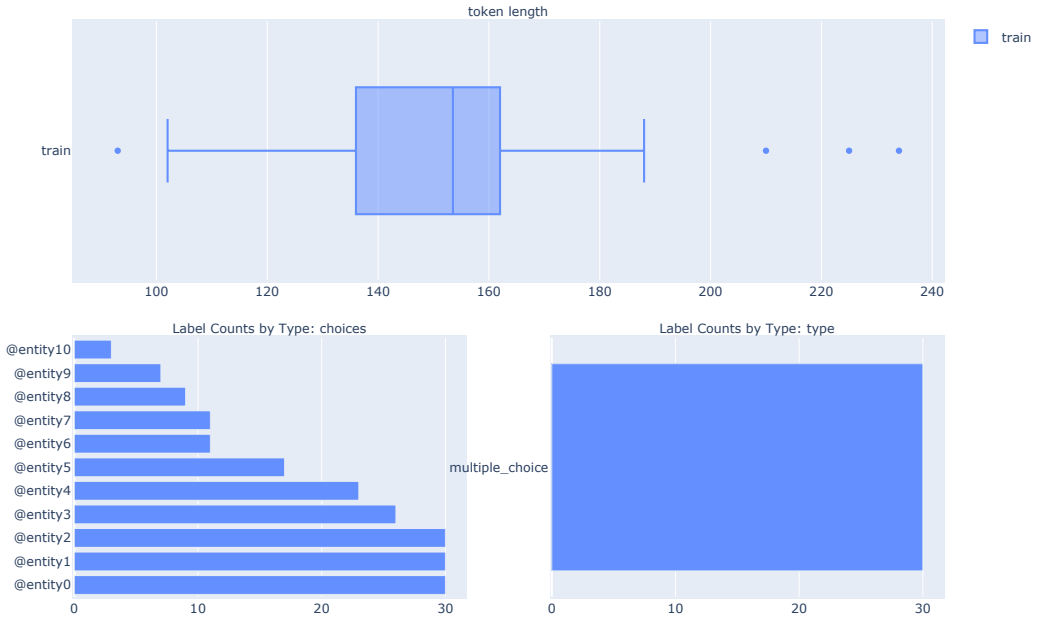

Figure 19: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We introduce BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments show that simple heuristics do not perform well on the new dataset and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Non-expert human performance is also higher on the new dataset compared to BIOREAD, and biomedical experts perform even better. We also introduce a new BERT-based MRC model, the best version of which substantially outperforms all other methods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make the new dataset available in three different sizes, also releasing our code, and providing a leader board.

Homepage: `https://github.com/PetrosStav/BioMRC_code`

URL: `https://github.com/PetrosStav/BioMRC_code`

Licensing: License information unavailable

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train

BioNLP 2009 Data Card

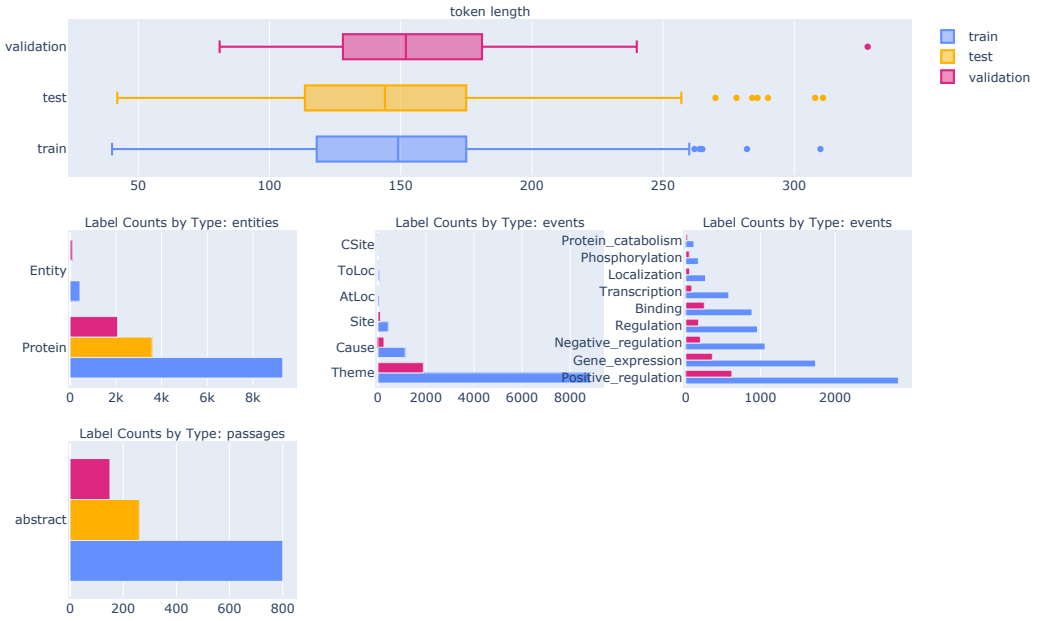

Figure 20: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioNLP Shared Task 2009 was organized by GENIA Project and its corpora were curated based on the annotations of the publicly available GENIA Event corpus and an unreleased (blind) section of the GENIA Event corpus annotations, used for evaluation. **Homepage:** `http://www.geniaproject.org/shared-tasks/bionlp-shared-task-2009`

URL: `http://www.geniaproject.org/shared-tasks/bionlp-shared-task-2009`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: coreference resolution, named entity recognition, event extraction

Schemas: `KB`

Splits: train, test, validation

BioNLP 2011 EPI Data Card

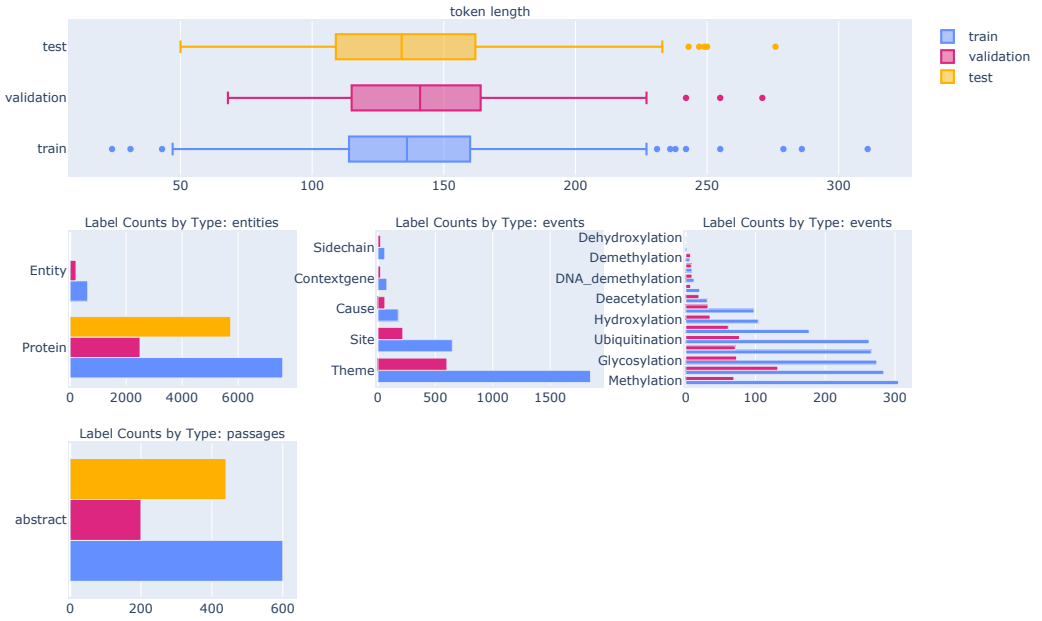

Figure 21: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The dataset of the Epigenetics and Post-translational Modifications (EPI) taskof BioNLP Shared Task 2011.

Homepage: `https://github.com/openbiocorpora/bionlp-st-2011-epi`

URL: `https://github.com/openbiocorpora/bionlp-st-2011-epi`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: event extraction, coreference resolution, named entity recognition

Schemas: `KB`

Splits: train, validation, test

BioNLP 2011 GE Data Card

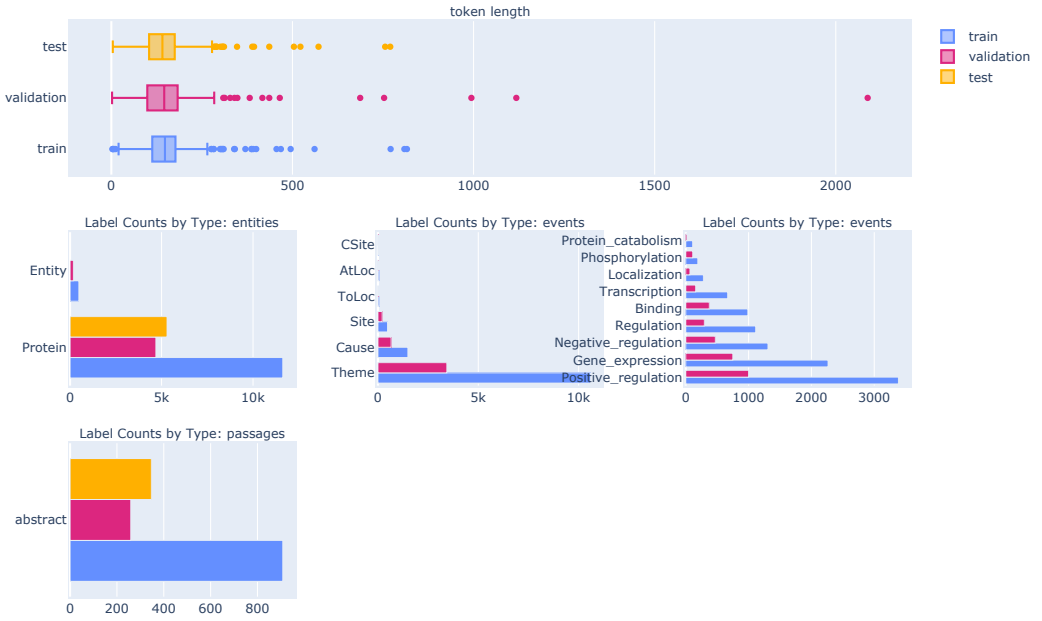

Figure 22: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioNLP-ST GE task has been promoting development of fine-grained information extraction (IE) from biomedical documents, since 2009. Particularly, it has focused on the domain of NFkB as a model domain of Biomedical IE.The GENIA task aims at extracting events occurring upon genes or gene products, which are typed as "Protein"without differentiating genes from gene products. Other types of physical entities, e.g. cells, cell components,are not differentiated from each other, and their type is given as "Entity".

Homepage: `https://sites.google.com/site/bionlpst/bionlp-shared-task-2011/genia-event-extraction-genia`

URL: `https://sites.google.com/site/bionlpst/bionlp-shared-task-2011/genia-event-extraction-genia`

Licensing: Creative Commons Attribution 3.0 Unported

Languages: English

Tasks: coreference resolution, event extraction, named entity recognition

Schemas: `KB`

Splits: train, validation, test

BioNLP 2011 ID Data Card

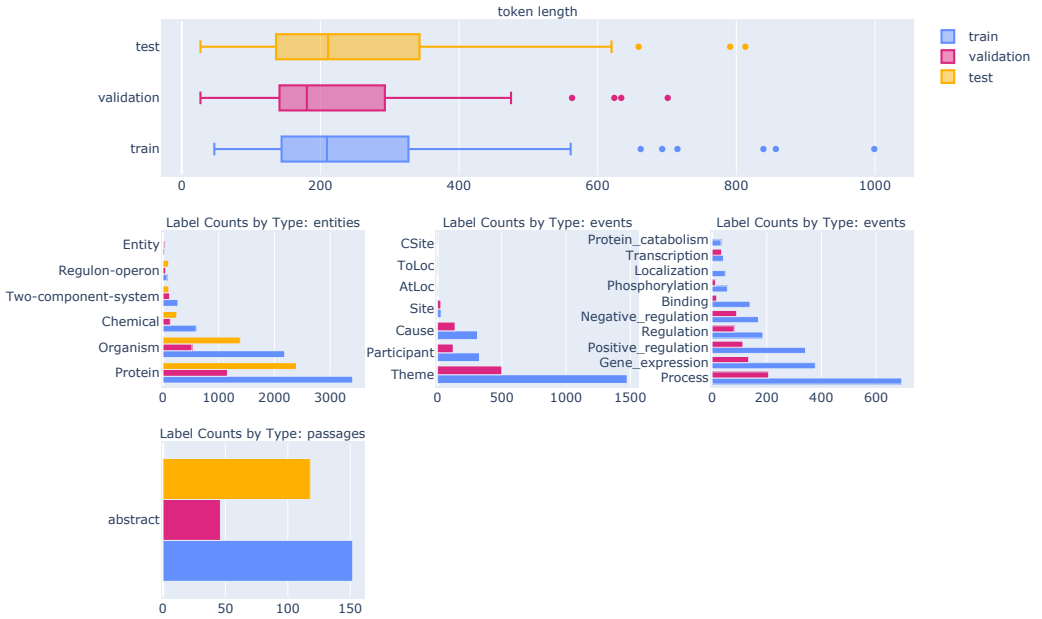

Figure 23: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The dataset of the Infectious Diseases (ID) task ofBioNLP Shared Task 2011.

Homepage: `https://github.com/openbiocorpora/bionlp-st-2011-id`

URL: `https://github.com/openbiocorpora/bionlp-st-2011-id`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: named entity recognition, event extraction, coreference resolution

Schemas: `KB`

Splits: train, validation, test

BioNLP 2011 REL Data Card

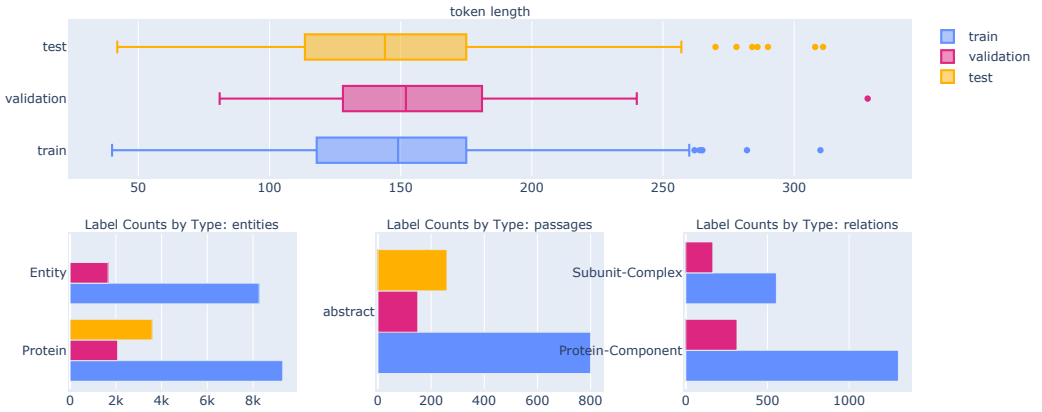

Figure 24: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The Entity Relations (REL) task is a supporting task of the BioNLP Shared Task 2011.The task concerns the extraction of two types of part-of relations between agene/protein and an associated entity. **Homepage:** `https://github.com/openbiocorpora/bionlp-st-2011-rel`

URL: `https://github.com/openbiocorpora/bionlp-st-2011-rel`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: named entity recognition, relation extraction, coreference resolution

Schemas: `KB`

Splits: train, validation, test

BioNLP 2013 CG Data Card

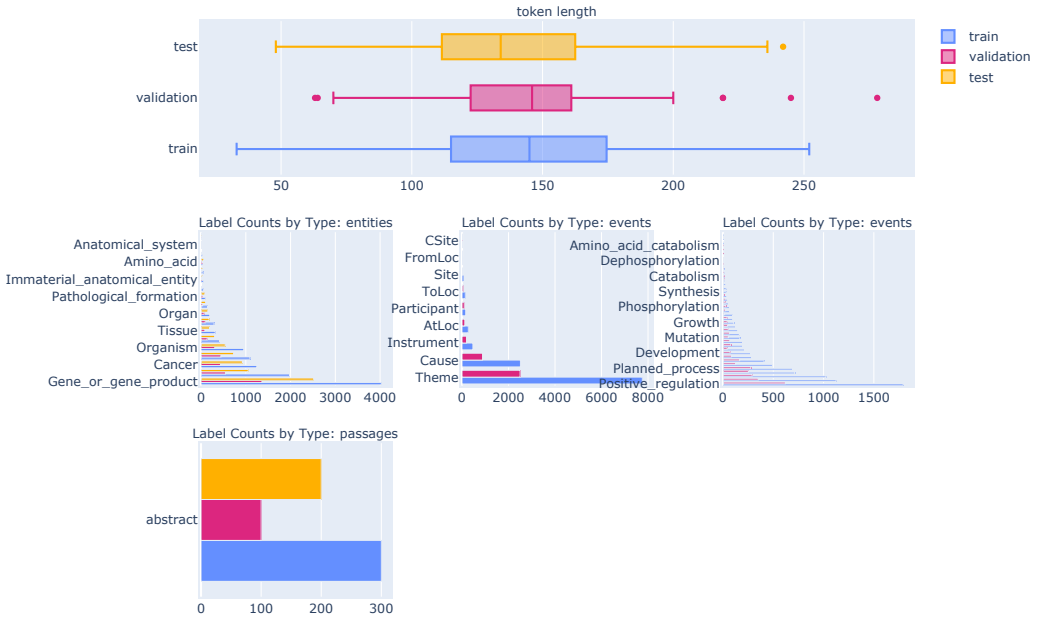

Figure 25: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: the Cancer Genetics (CG) is a event extraction task and a main task of the BioNLP Shared Task (ST) 2013.The CG task is an information extraction task targeting the recognition of events in text,represented as structured n-ary associations of given physical entities. In addition to addressing the cancer domain, the CG task is differentiated from previous event extraction tasks in the BioNLP ST series in addressing a wide range of pathological processes and multiple levels of biological organization, ranging from the molecular through the cellular and organ levels up to whole organisms. Final test set submissions were accepted from six teams

Homepage: `https://github.com/openbiocorpora/bionlp-st-2013-cg`

URL: `https://github.com/openbiocorpora/bionlp-st-2013-cg`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: named entity recognition, event extraction, coreference resolution

Schemas: `KB`

Splits: train, validation, test

BioNLP 2013 GE Data Card

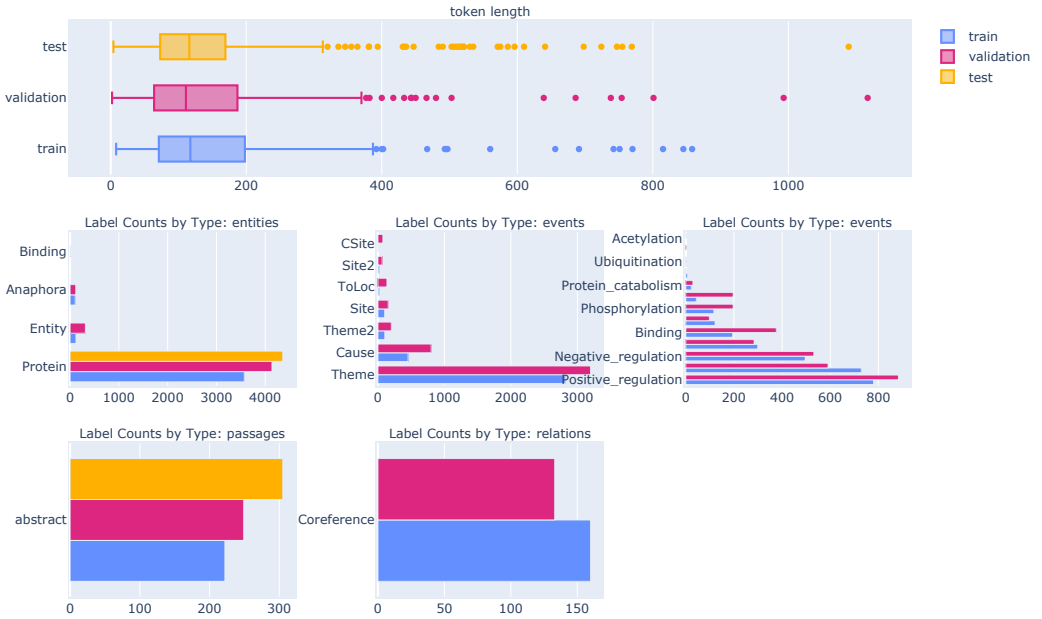

Figure 26: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioNLP-ST GE task has been promoting development of fine-grained information extraction (IE) from biomedical documents, since 2009. Particularly, it has focused on the domain of NFkB as a model domain of Biomedical IE

Homepage: `https://github.com/openbiocorpora/bionlp-st-2013-ge`

URL: `https://github.com/openbiocorpora/bionlp-st-2013-ge`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: coreference resolution, event extraction, named entity recognition, relation extraction

Schemas: `KB`

Splits: train, validation, test

BioNLP 2013 GRO Data Card

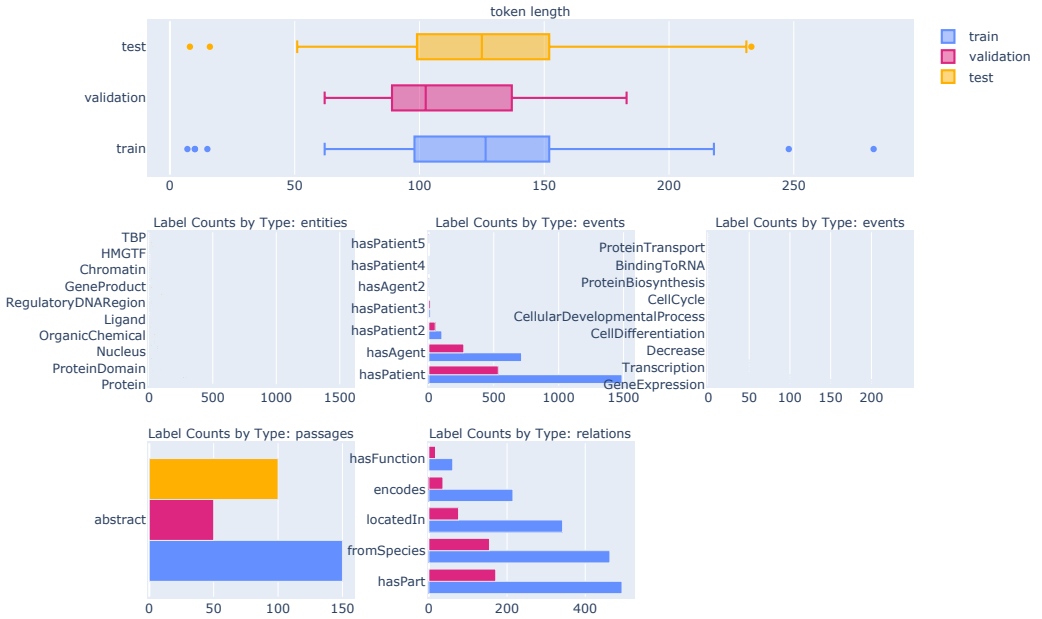

Figure 27: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: GRO Task: Populating the Gene Regulation Ontology with events andrelations. A data set from the bio NLP shared tasks competition from 2013.

Homepage: `https://github.com/openbiocorpora/bionlp-st-2013-gro`

URL: `https://github.com/openbiocorpora/bionlp-st-2013-gro`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: event extraction, relation extraction, named entity recognition

Schemas: `KB`

Splits: train, validation, test

BioNLP 2013 PC Data Card

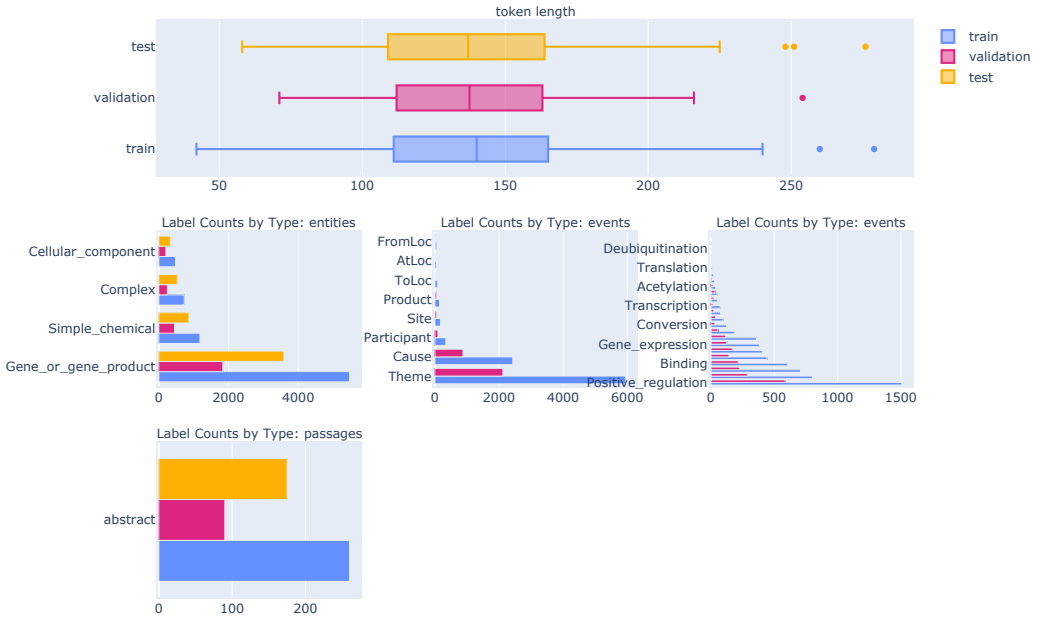

Figure 28: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: the Pathway Curation (PC) task is a main event extraction task of the BioNLP shared task (ST) 2013.The PC task concerns the automatic extraction of biomolecular reactions from text.The task setting, representation and semantics are defined with respect to pathway model standards and ontologies (SBML, BioPAX, SBO) and documents selected by relevance to specific model reactions. Two BioNLP ST 2013 participants successfully completed the PC task. The highest achieved F-score, 52.8**Homepage:** `https://github.com/openbiocorpora/bionlp-st-2013-pc`

URL: `https://github.com/openbiocorpora/bionlp-st-2013-pc`

Licensing: GENIA Project License for Annotated Corpora

Languages: English

Tasks: coreference resolution, named entity recognition, event extraction

Schemas: `KB`

Splits: train, validation, test

BioNLP 2019 BB Data Card

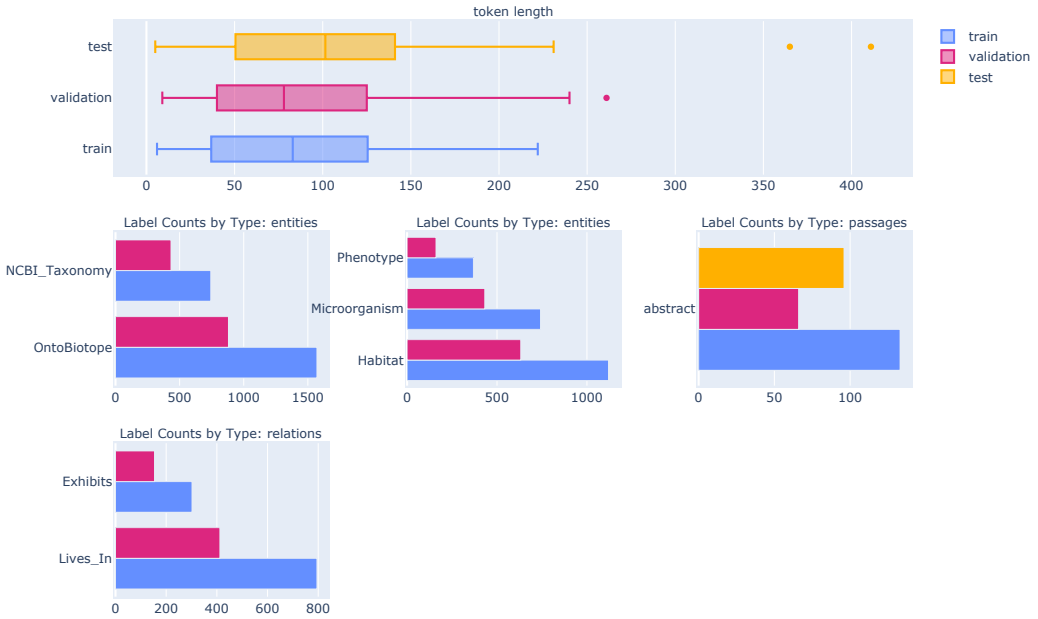

Figure 29: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The task focuses on the extraction of the locations and phenotypes of microorganisms from PubMed abstracts and full-text excerpts, and the characterization of these entities with respect to reference knowledge sources (NCBI taxonomy, OntoBiotope ontology). The task is motivated by the importance of the knowledge on biodiversity for fundamental research and applications in microbiology.

Homepage: `https://sites.google.com/view/bb-2019/dataset`

URL: `https://sites.google.com/view/bb-2019/dataset`

Licensing: License information unavailable

Languages: English

Tasks: named entity disambiguation, relation extraction, named entity recognition

Schemas: `KB`

Splits: train, validation, test

BioRED Data Card

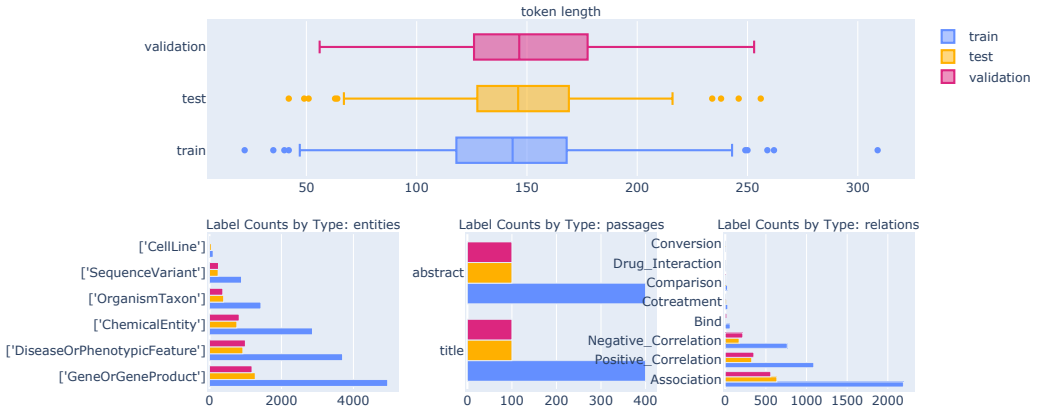

Figure 30: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Relation Extraction corpus with multiple entity types (e.g., gene/protein, disease, chemical) and relation pairs (e.g., gene-disease; chemical-chemical), on a set of 600 PubMed articles

Homepage: `https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/`

URL: `https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/`

Licensing: License information unavailable

Languages: English

Tasks: named entity recognition, relation extraction

Schemas: `KB`

Splits: train, test, validation

BioRelEx Data Card

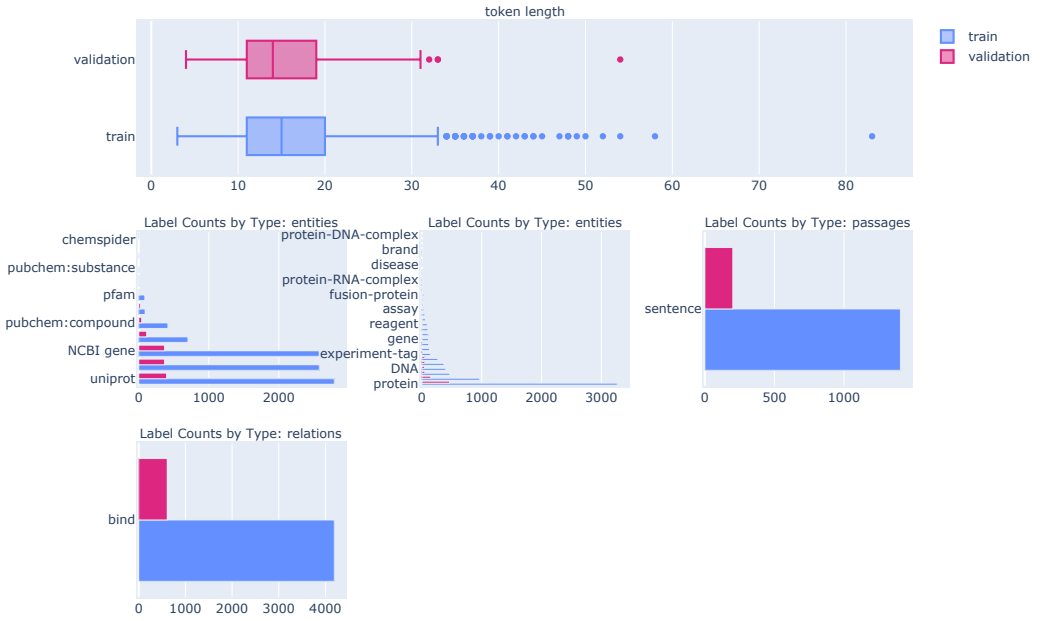

Figure 31: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: BioRelEx is a biological relation extraction dataset. Version 1.0 contains 2010 annotated sentences that describe binding interactions between various biological entities (proteins, chemicals, etc.). 1405 sentences are for training,another 201 sentences are for validation. They are publicly available at `https://github.com/YerevaNN/BioRelEx/releases`. Another 404 sentences are for testing which are kept private for at this Codalab competition `https://competitions.codalab.org/competitions/20468`. All sentences contain words "bind", "bound" or "binding". For every sentence we provide: 1) Complete annotations of all biological entities that appear in the sentence 2) Entity types (32types) and grounding information for most of the proteins and families (links to uniprot, interpro and other databases)3) Coreference between entities in the same sentence (e.g. abbreviations and synonyms) 4) Binding interactions between the annotated entities 5) Binding interaction types: positive, negative (A does not bind B) and neutral (A may bind to B)

Homepage: `https://github.com/YerevaNN/BioRelEx`

URL: `https://github.com/YerevaNN/BioRelEx`

Licensing: License information unavailable

Languages: English

Tasks: named entity recognition, coreference resolution, named entity disambiguation, relation extraction

Schemas: `KB`

Splits: train, validation

BioScope Data Card

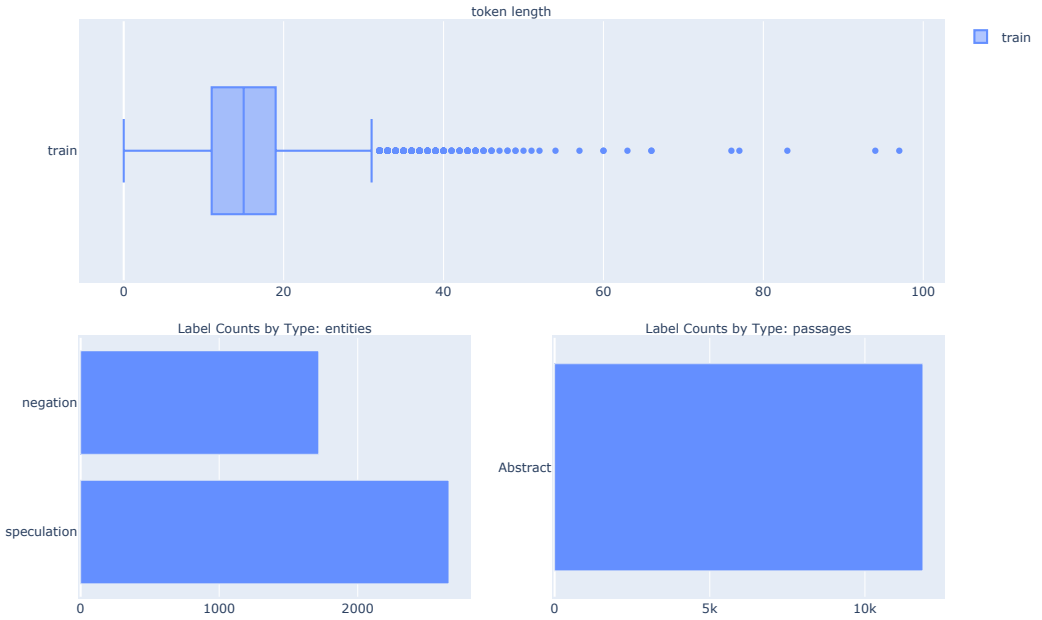

Figure 32: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioScope corpus consists of medical and biological texts annotated for negation, speculation and their linguistic scope. This was done to allow a comparison between the development of systems for negation/hedge detection and scope resolution. The BioScope corpus was annotated by two independent linguists following the guidelines written by our linguist expert before the annotation of the corpus was initiated.

Homepage: `https://rgai.inf.u-szeged.hu/node/105`

URL: `https://rgai.inf.u-szeged.hu/node/105`

Licensing: Creative Commons Attribution 2.0 Generic

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

BioScope Data Card

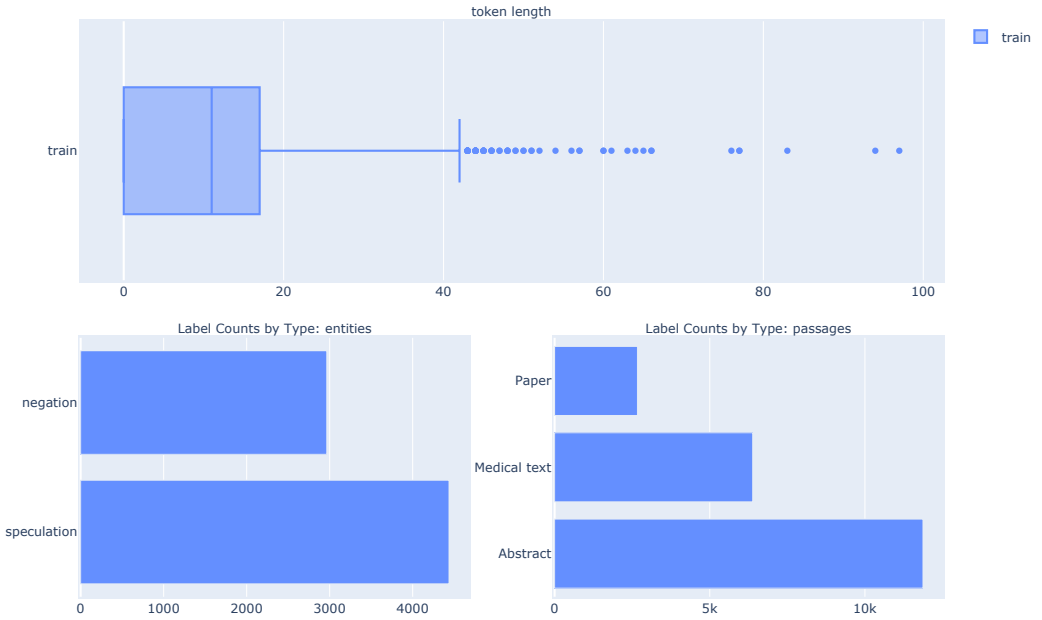

Figure 33: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioScope corpus consists of medical and biological texts annotated for negation, speculation and their linguistic scope. This was done to allow a comparison between the development of systems for negation/hedge detection and scope resolution. The BioScope corpus was annotated by two independent linguists following the guidelines written by our linguist expert before the annotation of the corpus was initiated.

Homepage: `https://rgai.inf.u-szeged.hu/node/105`

URL: `https://rgai.inf.u-szeged.hu/node/105`

Licensing: Creative Commons Attribution 2.0 Generic

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

BioScope (Medical Texts) Data Card

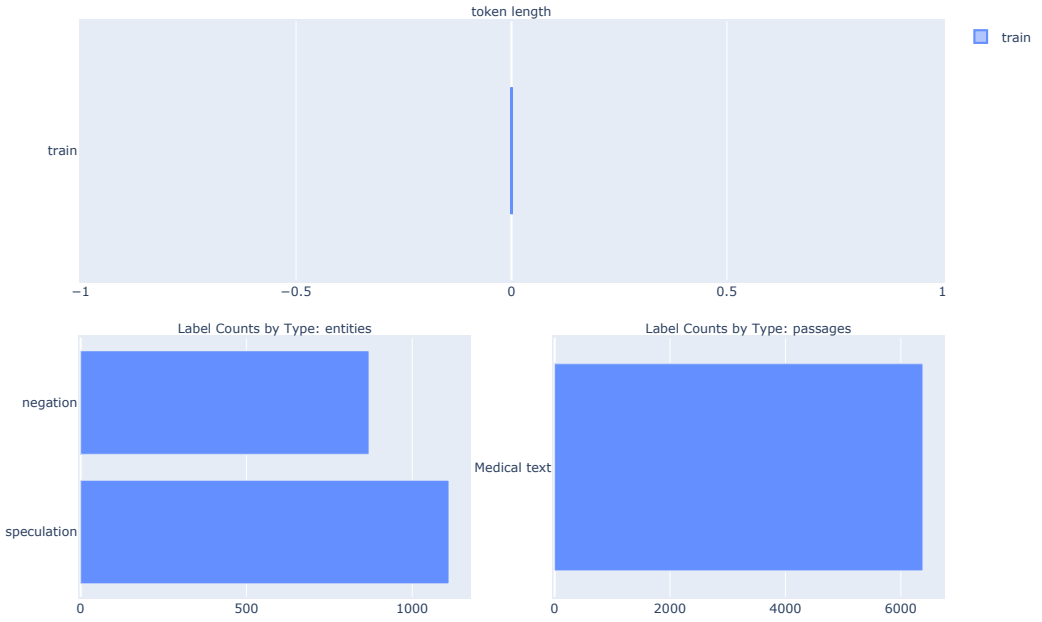

Figure 34: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description The BioScope corpus consists of medical and biological texts annotated for negation, speculation and their linguistic scope. This was done to allow a comparison between the development of systems for negation/hedge detection and scope resolution. The BioScope corpus was annotated by two independent linguists following the guidelines written by our linguist expert before the annotation of the corpus was initiated.

Homepage: `https://rgai.inf.u-szeged.hu/node/105`

URL: `https://rgai.inf.u-szeged.hu/node/105`

Licensing: Creative Commons Attribution 2.0 Generic

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

BioScope (Papers) Data Card

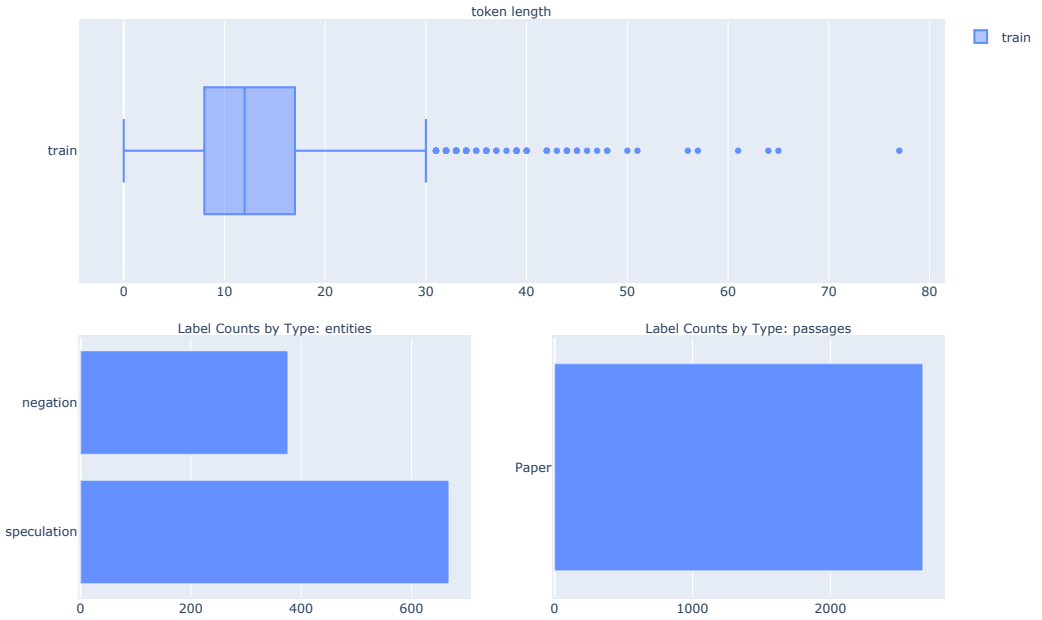

Figure 35: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioScope corpus consists of medical and biological texts annotated for negation, speculation and their linguistic scope. This was done to allow a comparison between the development of systems for negation/hedge detection and scope resolution. The BioScope corpus was annotated by two independent linguists following the guidelines written by our linguist expert before the annotation of the corpus was initiated.

Homepage: `https://rgai.inf.u-szeged.hu/node/105`

URL: `https://rgai.inf.u-szeged.hu/node/105`

Licensing: Creative Commons Attribution 2.0 Generic

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

BIOSSES Data Card

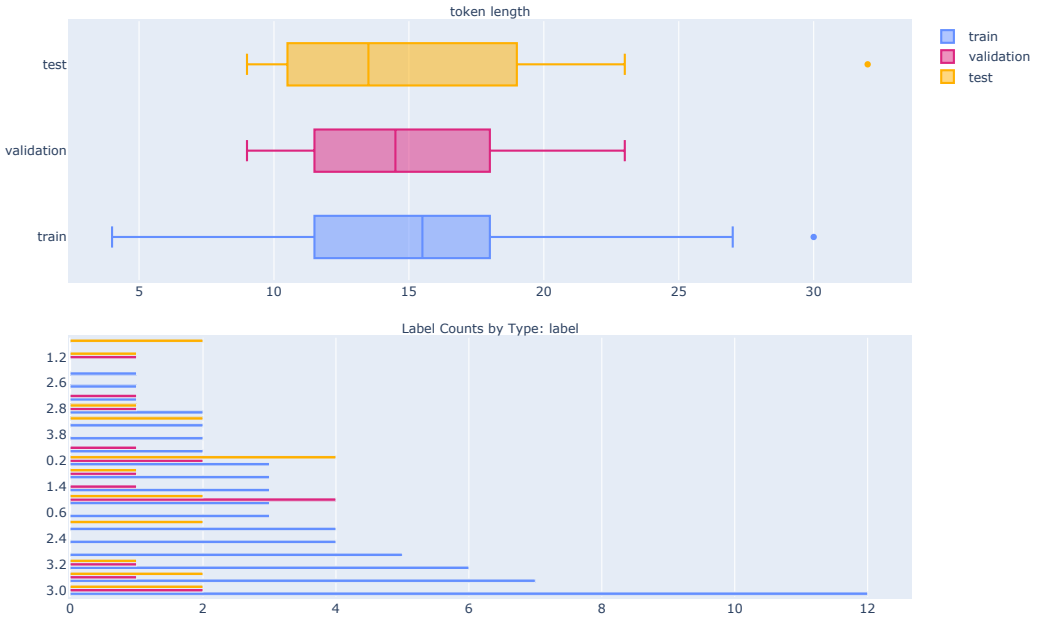

Figure 36: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: BioSSES computes similarity of biomedical sentences by utilizing WordNet as the general domain ontology and UMLS as the biomedical domain specific ontology.The original paper outlines the approaches with respect to using annotator score as golden standard. Source view will return all annotator score individually whereas the Bigbio view will return the mean of the annotator score.

Homepage: `https://tabilab.cmpe.boun.edu.tr/BIOSSES/DataSet.html`

URL: `https://tabilab.cmpe.boun.edu.tr/BIOSSES/DataSet.html`

Licensing: GNU General Public License v3.0 only

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train, validation, test

CADEC Data Card

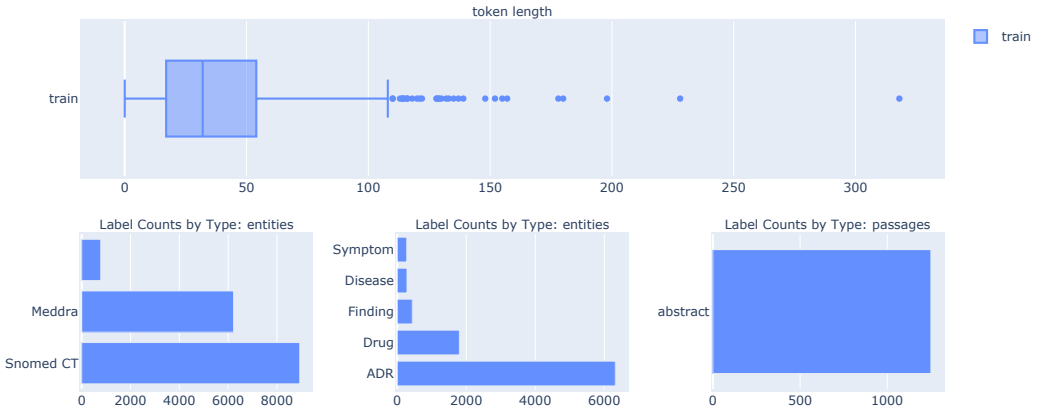

Figure 37: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The CADEC corpus (CSIRO Adverse Drug Event Corpus) is is a new rich annotated corpus of medical forum posts on patient-reported Adverse Drug Events (ADEs). The corpus is sourced from posts on social media, and contains text that is largely written in colloquial language and often deviates from formal English grammar and punctuation rules.Annotations contain mentions of concepts such as drugs, adverse events, symptoms, and diseases linked to their corresponding concepts in controlled vocabularies, i.e., SNOMED Clinical Terms and MedDRA. The quality of the annotations is ensured by annotation guidelines, multi-stage annotations, measuring inter-annotator agreement, and final review of the annotations by a clinical terminologist. This corpus is useful for those studies in the area of information extraction, or more generally text mining, from social media to detect possible adverse drug reactions from direct patient reports. The dataset contains three views: original (entities annotated in the posts), meddra (entities normalized with meddra codes), sct (entities normalized with SNOMED CT codes).

Homepage: `https://data.gov.au/dataset/ds-dap-csiro%3A10948/details?q=`

URL: `https://data.gov.au/dataset/ds-dap-csiro%3A10948/details?q=`

Licensing: CSIRO Data License

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train

CANTEMIST Data Card

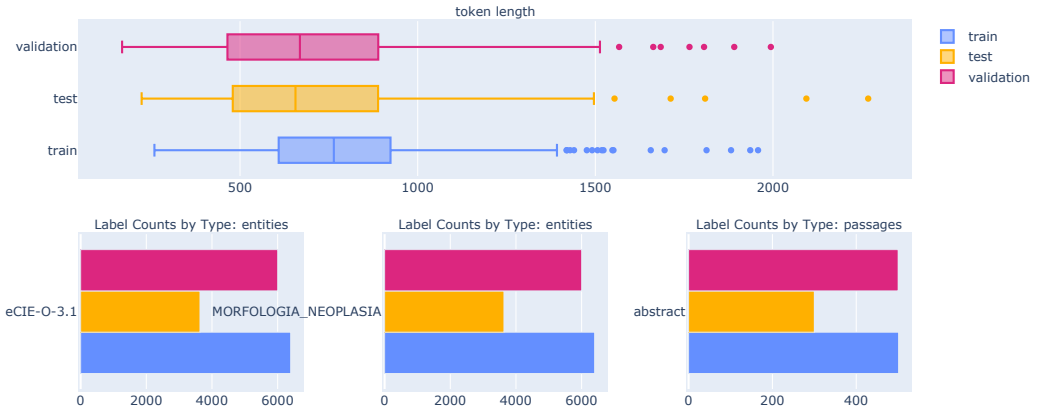

Figure 38: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Collection of 1301 oncological clinical case reports written in Spanish, with tumor morphology mentions manually annotated and mapped by clinical experts to a controlled terminology. Every tumor morphology mention is linked to an eCIE-O code (the Spanish equivalent of ICD-O).The original dataset is distributed in Brat format, and was randomly sampled into 3 subsets. The training, development and test sets contain 501, 500 and 300 documents each, respectively.This dataset was designed for the CANcer TExt Mining Shared Task, sponsored by Plan-TL. The task is divided in 3 subtasks: CANTEMIST-NER, CANTEMIST_NORM and CANTEMIST-CODING.

CANTEMIST-NER track: requires finding automatically tumor morphology mentions. All tumor morphology mentions are defined by their corresponding character offsets in UTF-8 plain text medical documents.

CANTEMIST-NORM track: clinical concept normalization or named entity normalization task that requires to return all tumor morphology entity mentions together with their corresponding eCIE-O-3.1 codes i.e. finding and normalizing tumor morphology mentions.

CANTEMIST-CODING track: requires returning for each of document a ranked list of its corresponding ICD-O-3 codes. This it is essentially a sort of indexing or multi-label classification task or oncology clinical coding.

For further information, please visit `https://temu.bsc.es/cantemist` or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/cantemist/?p=4338`

URL: `https://temu.bsc.es/cantemist/?p=4338`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train, test, validation

Cantemist Data Card

Figure 39: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description Collection of 1301 oncological clinical case reports written in Spanish, with tumor morphology mentions manually annotated and mapped by clinical experts to a controlled terminology. Every tumor morphology mention is linked to an eCIE-O code (the Spanish equivalent of ICD-O).The original dataset is distributed in Brat format, and was randomly sampled into 3 subsets. The training, development and test sets contain 501, 500 and 300 documents each, respectively.This dataset was designed for the CANcer TExt Mining Shared Task, sponsored by Plan-TL. The task is divided in 3 subtasks: CANTEMIST-NER, CANTEMIST_NORM and CANTEMIST-CODING.

CANTEMIST-NER track: requires finding automatically tumor morphology mentions. All tumor morphology mentions are defined by their corresponding character offsets in UTF-8 plain text medical documents.

CANTEMIST-NORM track: clinical concept normalization or named entity normalization task that requires to return all tumor morphology entity mentions together with their corresponding eCIE-O-3.1 codes i.e. finding and normalizing tumor morphology mentions.

CANTEMIST-CODING track: requires returning for each of document a ranked list of its corresponding ICD-O-3 codes. This it is essentially a sort of indexing or multi-label classification task or oncology clinical coding.

For further information, please visit `https://temu.bsc.es/cantemist` or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/cantemist/?p=4338`

URL: `https://temu.bsc.es/cantemist/?p=4338`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `KB TEXT`

Splits: train, test, validation

CellFinder Data Card

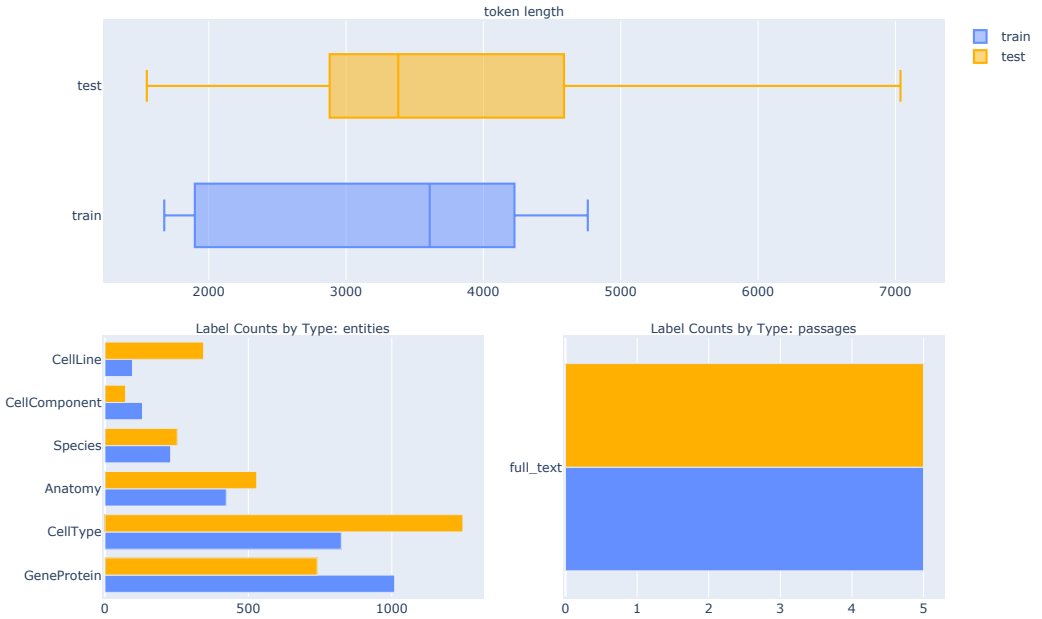

Figure 40: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The CellFinder project aims to create a stem cell data repository by linking information from existing public databases and by performing text mining on the research literature. The first version of the corpus is composed of 10 full text documents containing more than 2,100 sentences, 65,000 tokens and 5,200 annotations for entities. The corpus has been annotated with six types of entities (anatomical parts, cell components, cell lines, cell types, genes/protein and species) with an overall inter-annotator agreement around 80%.(see `https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/cellfinder/`).

Homepage: `https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/cellfinder/`

URL: `https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/cellfinder/`

Licensing: Creative Commons Attribution Share Alike 3.0 Unported

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test

cellfinder Data Card

Figure 41: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description The CellFinder project aims to create a stem cell data repository by linking information from existing public databases and by performing text mining on the research literature. The first version of the corpus is composed of 10 full text documents containing more than 2,100 sentences, 65,000 tokens and 5,200 annotations for entities. The corpus has been annotated with six types of entities (anatomical parts, cell components, cell lines, cell types, genes/protein and species) with an overall inter-annotator agreement around 80%.(see https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/cellfinder/).

Homepage: `https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/cellfinder/`

URL: `https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/cellfinder/`

Licensing: Creative Commons Attribution Share Alike 3.0 Unported

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test

CHEBI Corpus (Abstracts 1) Data Card

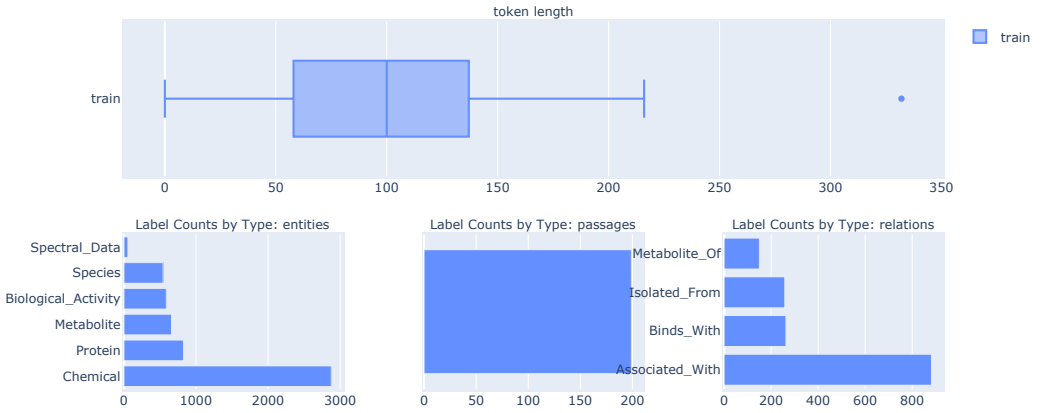

Figure 42: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The ChEBI corpus contains 199 annotated abstracts and 100 annotated full papers. All documents in the corpus have been annotated for named entities and relations between these.In total, our corpus provides over 15000 named entity annotations and over 6,000 relations between entities.

Homepage: `http://www.nactem.ac.uk/chebi`

URL: `http://www.nactem.ac.uk/chebi`

Licensing: Creative Commons Attribution 4.0 International

Languages: English

Tasks: relation extraction, named entity recognition

Schemas: `KB`

Splits: train

CHEBI Corpus (Abstracts 2) Data Card

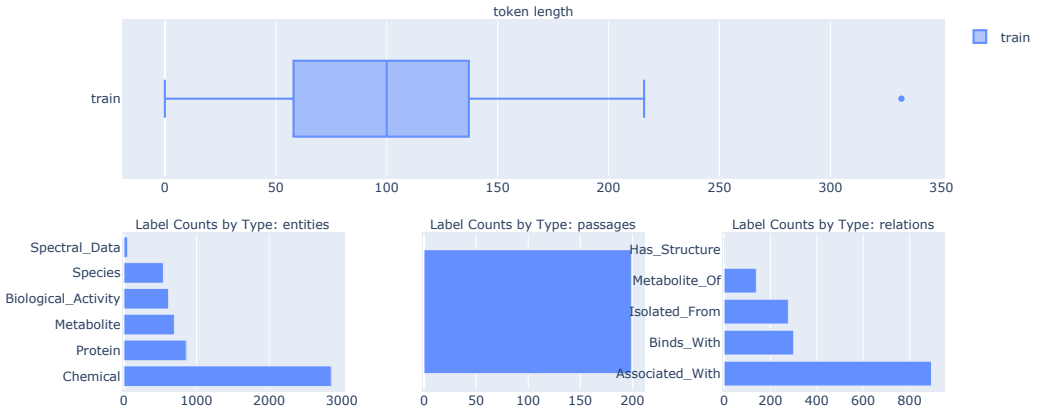

Figure 43: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The ChEBI corpus contains 199 annotated abstracts and 100 annotated full papers.All documents in the corpus have been annotated for named entities and relations between these.In total, our corpus provides over 15000 named entity annotations and over 6,000 relations between entities.

Homepage: `http://www.nactem.ac.uk/chebi`

URL: `http://www.nactem.ac.uk/chebi`

Licensing: Creative Commons Attribution 4.0 International

Languages: English

Tasks: relation extraction, named entity recognition

Schemas: `KB`

Splits: train

CHEBI (Papers) Data Card

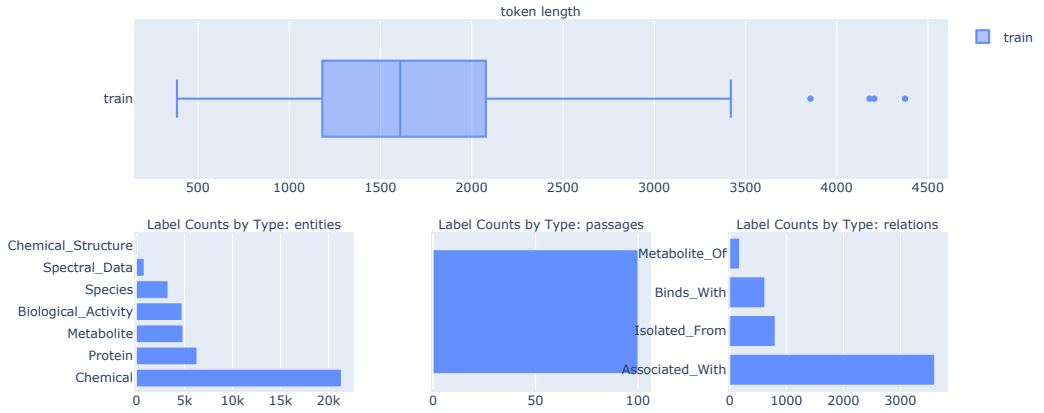

Figure 44: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The ChEBI corpus contains 199 annotated abstracts and 100 annotated full papers.All documents in the corpus have been annotated for named entities and relations between these.In total, our corpus provides over 15000 named entity annotations and over 6,000 relations between entities.

Homepage: `http://www.nactem.ac.uk/chebi`

URL: `http://www.nactem.ac.uk/chebi`

Licensing: Creative Commons Attribution 4.0 International

Languages: English

Tasks: relation extraction, named entity recognition

Schemas: `KB`

Splits: train

CHEMDNER Data Card

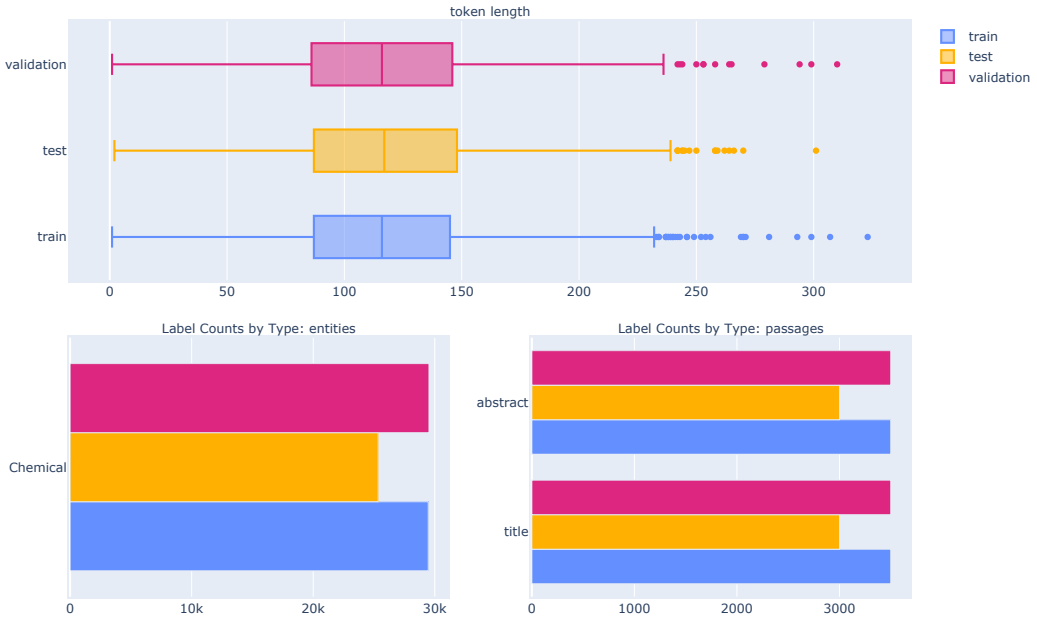

Figure 45: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We present the CHEMDNER corpus, a collection of 10,000 PubMed abstracts that contain a total of 84,355 chemical entity mentions labeled manually by expert chemistry literature curators, following annotation guidelines specifically defined for this task. The abstracts of the CHEMDNER corpus were selected to be representative for all major chemical disciplines. Each of the chemical entity mentions was manually labeled according to its structure-associated chemical entity mention (SACEM) class: abbreviation, family, formula, identifier, multiple, systematic and trivial.

Homepage: `https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/`

URL: `https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/`

Licensing: License information unavailable

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

CHEMDNER Data Card

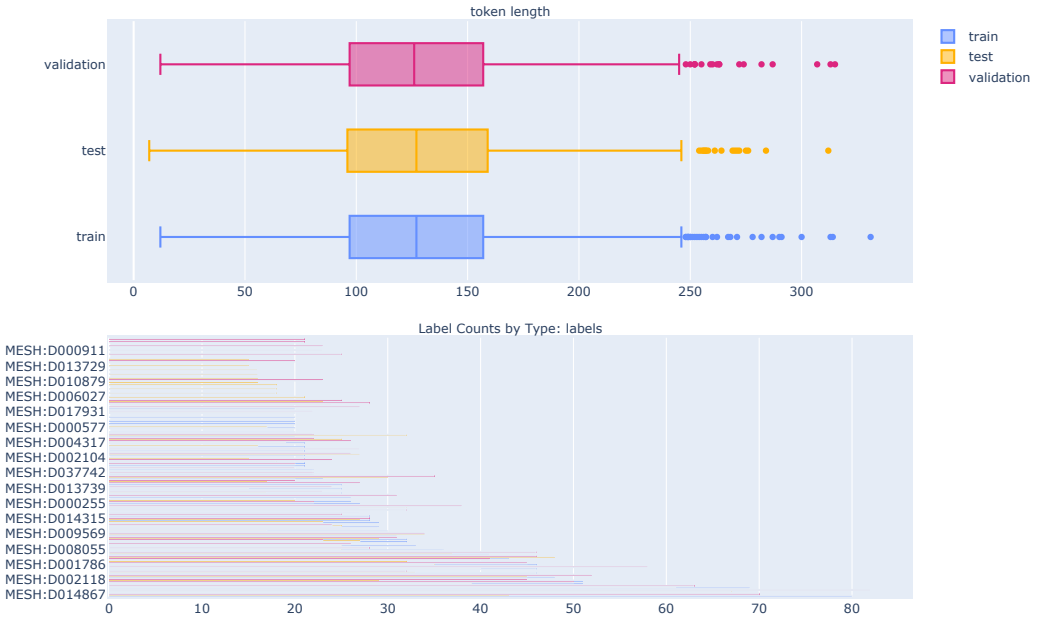

Figure 46: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We present the CHEMDNER corpus, a collection of 10,000 PubMed abstracts that contain a total of 84,355 chemical entity mentions labeled manually by expert chemistry literature curators, following annotation guidelines specifically defined for this task. The abstracts of the CHEMDNER corpus were selected to be representative for all major chemical disciplines. Each of the chemical entity mentions was manually labeled according to its structure-associated chemical entity mention (SACEM) class: abbreviation, family, formula, identifier, multiple, systematic and trivial.

Homepage: `https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/`

URL: `https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/`

Licensing: License information unavailable

Languages: English

Tasks: text classification

Schemas: `TEXT`

Splits: train, test, validation

ChemProt Data Card

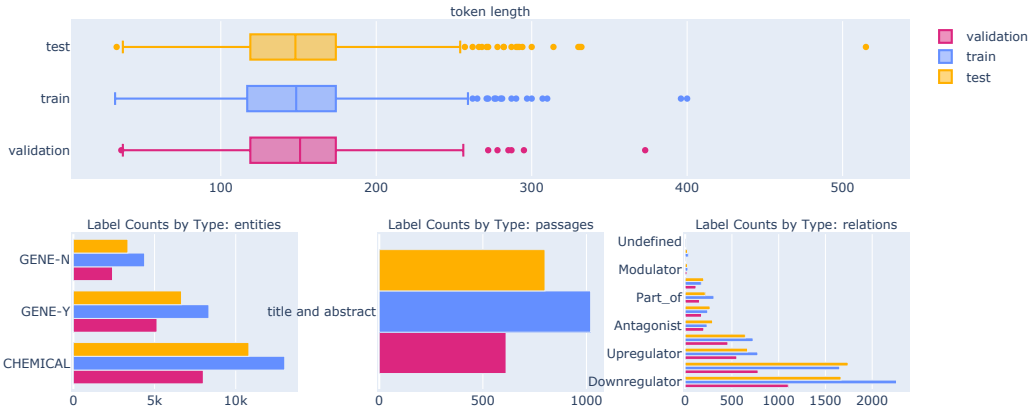

Figure 47: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The BioCreative VI Chemical-Protein interaction dataset identifies entities of chemicals and proteins and their likely relation to one other. Compounds are generally agonists (activators) or antagonists (inhibitors) of proteins.

Homepage: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-5/`

URL: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-5/`

Licensing: PUBLIC_DOMAIN_MARK_1p0

Languages: English

Tasks: named entity recognition, relation extraction

Schemas: `KB`

Splits: sample, train, test, validation

CHIA Data Card

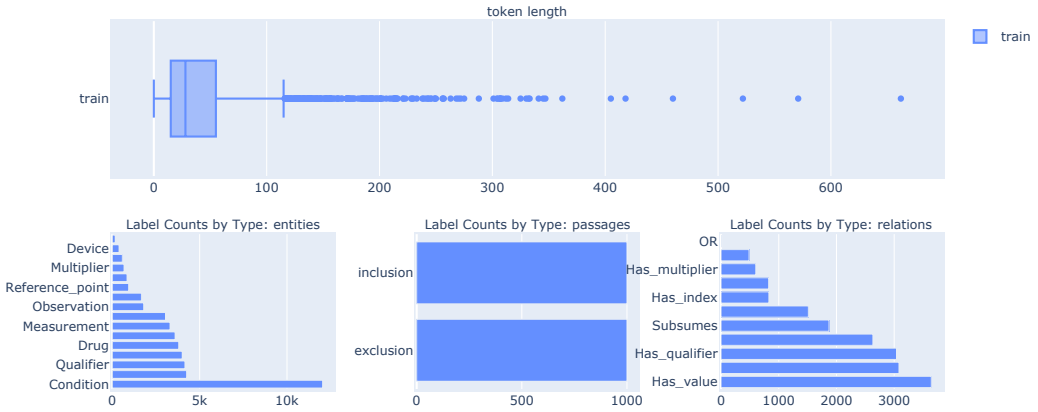

Figure 48: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: A large annotated corpus of patient eligibility criteria extracted from 1,000 interventional, Phase IV clinical trials registered in ClinicalTrials.gov. This dataset includes 12,409 annotated eligibility criteria, represented by 41,487 distinctive entities of 15 entity types and 25,017 relationships of 12 relationship types. **Homepage:** `https://github.com/WengLab-InformaticsResearch/CHIA`

URL: `https://github.com/WengLab-InformaticsResearch/CHIA`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition, relation extraction

Schemas: `KB`

Splits: train

citation gia test collection Data Card

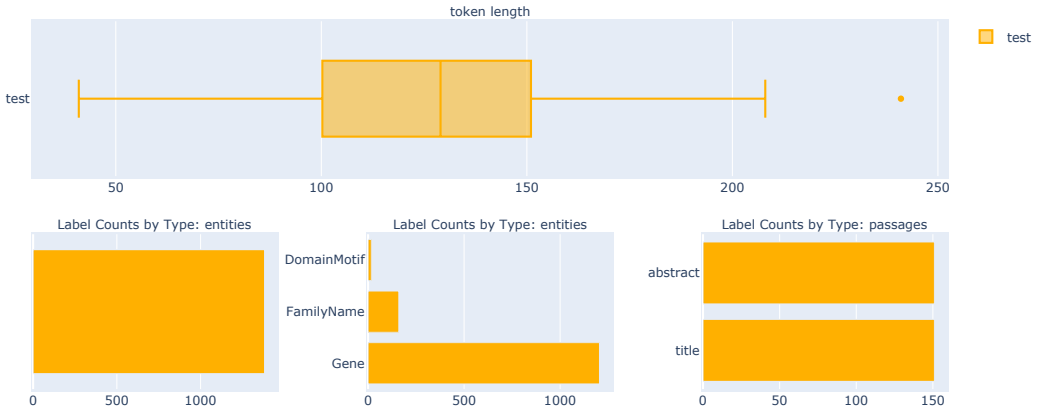

Figure 49: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The Citation GIA Test Collection was recently created for gene indexing at the NLM and includes 151 PubMed abstracts with both mention-level and document-level annotations. They are selected because both have a focus on human genes.

Homepage: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/gnormplus/`

URL: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/gnormplus/`

Licensing: Unknown

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: test

CodiEsp (Diagnosis Coding) Data Card

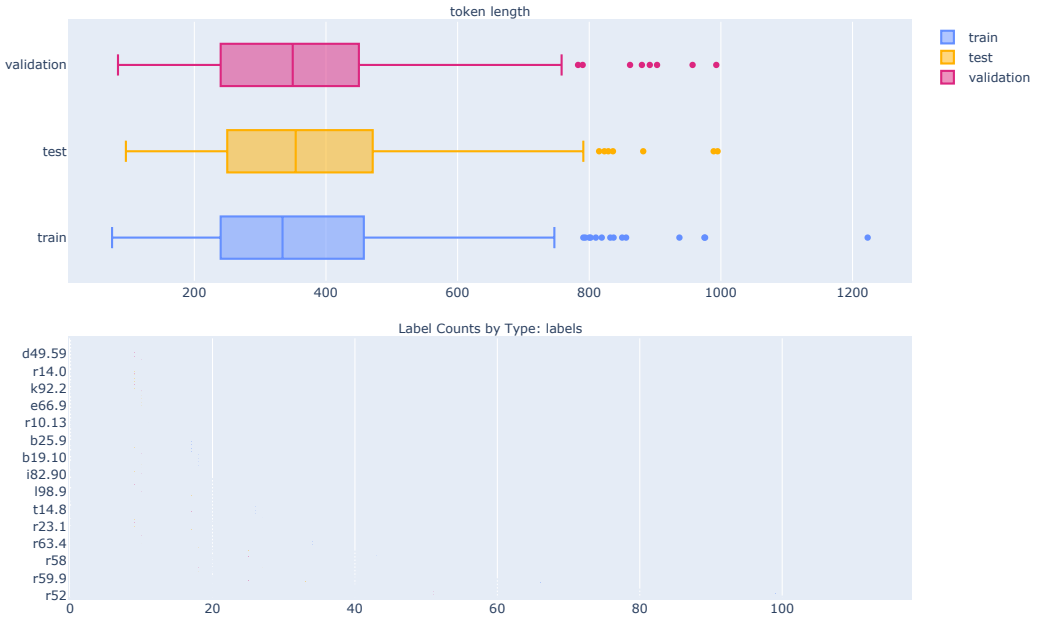

Figure 50: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Synthetic corpus of 1,000 manually selected clinical case studies in Spanish that was designed for the Clinical Case Coding in Spanish Shared Task, as part of the CLEF 2020 conference. The goal of the task was to automatically assign ICD10 codes (CIE-10, in Spanish) to clinical case documents, being evaluated against manually generated ICD10 codifications. The CodiEsp corpus was selected manually by practicing physicians and clinical documentalists and annotated by clinical coding professionals meeting strict quality criteria. They reached an inter-annotator agreement of 88.6% for diagnosis coding, 88.9% for procedure coding and 80.5% for the textual reference annotation.The final collection of 1,000 clinical cases that make up the corpus had a total of 16,504 sentences and 396,988 words. All documents are in Spanish language and CIE10 is the coding terminology (the Spanish version of ICD10-CM and ICD10-PCS). The CodiEsp corpus has been randomly sampled into three subsets. The train set contains 500 clinical cases, while the development and test sets have 250 clinical cases each. In addition to these, a collection of 176,294 abstracts from Lilacs and Ibecs with the corresponding ICD10 codes (ICD10-CM and ICD10-PCS) was provided by the task organizers. Every abstract has at least one associated code, with an average of 2.5 ICD10 codes per abstract.

The CodiEsp track was divided into three sub-tracks (2 main and 1 exploratory):

CodiEsp-D: The Diagnosis Coding sub-task, which requires automatic ICD10-CM [CIE10-Diagnóstico] code assignment.

CodiEsp-P: The Procedure Coding sub-task, which requires automatic ICD10-PCS [CIE10-Procedimiento] code assignment.

CodiEsp-X: The Explainable AI exploratory sub-task, which requires to submit the reference to the predicted codes (both ICD10-CM and ICD10-PCS). The goal of this novel task was not only to predict the correct codes but also to present the reference in the text that supports the code predictions.

For further information, please visit https://temu.bsc.es/codiesp or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/codiesp/`

URL: `https://temu.bsc.es/codiesp/`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `TEXT`

Splits: train, test, validation

CodiEsp (extra CIE) Data Card

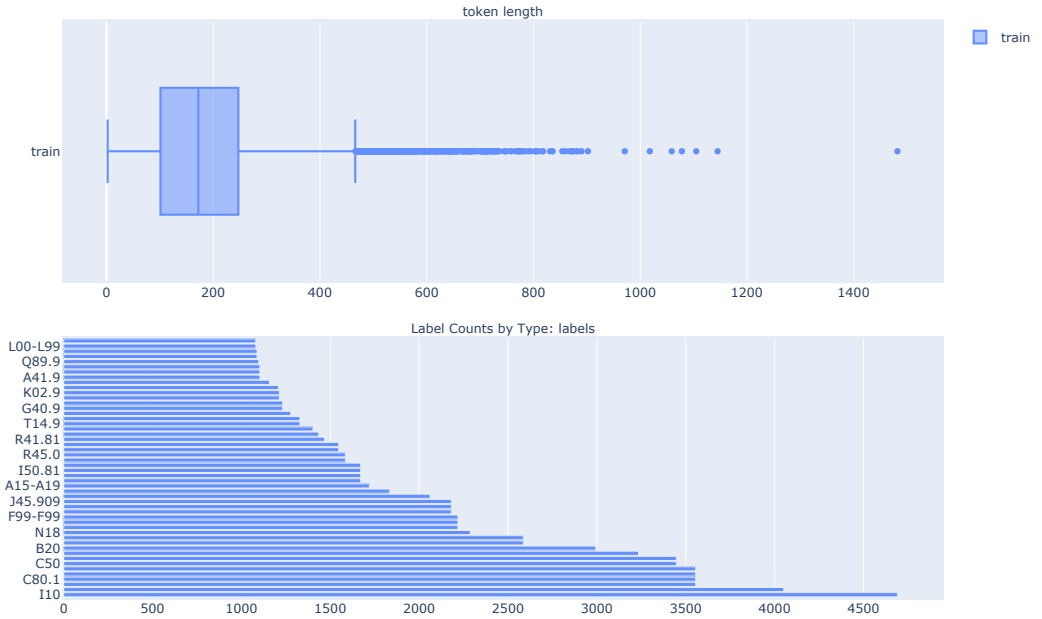

Figure 51: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Synthetic corpus of 1,000 manually selected clinical case studies in Spanish that was designed for the Clinical Case Coding in Spanish Shared Task, as part of the CLEF 2020 conference. The goal of the task was to automatically assign ICD10 codes (CIE-10, in Spanish) to clinical case documents, being evaluated against manually generated ICD10 codifications. The CodiEsp corpus was selected manually by practicing physicians and clinical documentalists and annotated by clinical coding professionals meeting strict quality criteria. They reached an inter-annotator agreement of 88.6% for diagnosis coding, 88.9% for procedure coding and 80.5% for the textual reference annotation.The final collection of 1,000 clinical cases that make up the corpus had a total of 16,504 sentences and 396,988 words. All documents are in Spanish language and CIE10 is the coding terminology (the Spanish version of ICD10-CM and ICD10-PCS). The CodiEsp corpus has been randomly sampled into three subsets. The train set contains 500 clinical cases, while the development and test sets have 250 clinical cases each. In addition to these, a collection of 176,294 abstracts from Lilacs and Ibecs with the corresponding ICD10 codes (ICD10-CM and ICD10-PCS) was provided by the task organizers. Every abstract has at least one associated code, with an average of 2.5 ICD10 codes per abstract.

The CodiEsp track was divided into three sub-tracks (2 main and 1 exploratory):

CodiEsp-D: The Diagnosis Coding sub-task, which requires automatic ICD10-CM [CIE10-Diagnóstico] code assignment.

CodiEsp-P: The Procedure Coding sub-task, which requires automatic ICD10-PCS [CIE10-Procedimiento] code assignment.

CodiEsp-X: The Explainable AI exploratory sub-task, which requires to submit the reference to the predicted codes (both ICD10-CM and ICD10-PCS). The goal of this novel task was not only to predict the correct codes but also to present the reference in the text that supports the code predictions.

For further information, please visit https://temu.bsc.es/codiesp or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/codiesp/`

URL: `https://temu.bsc.es/codiesp/`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `KB TEXT`

Splits: train

CodiEsp (extra mesh) Data Card

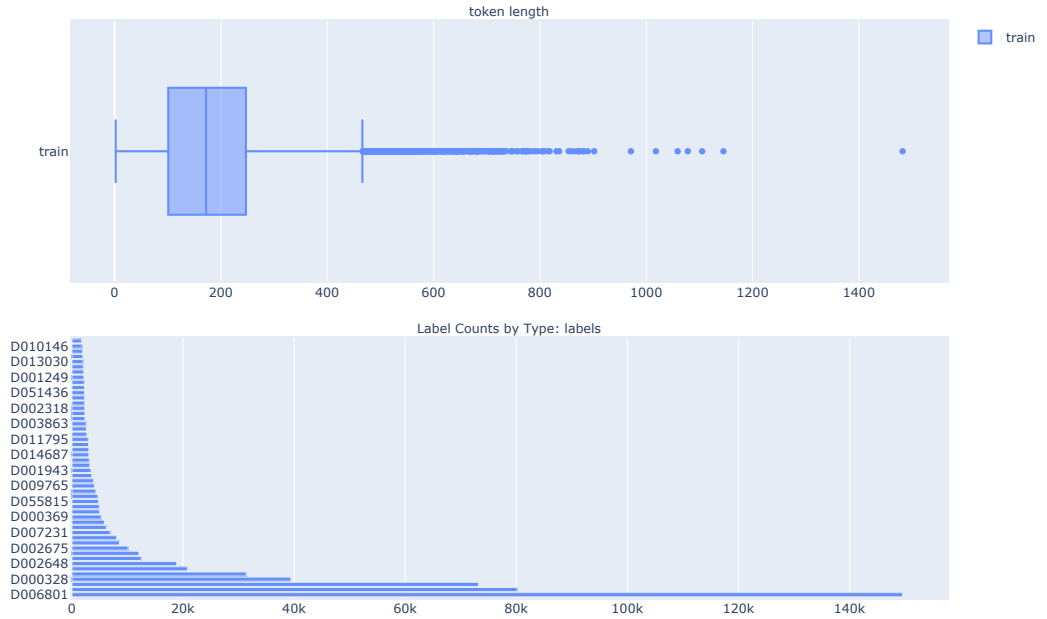

Figure 52: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Synthetic corpus of 1,000 manually selected clinical case studies in Spanish that was designed for the Clinical Case Coding in Spanish Shared Task, as part of the CLEF 2020 conference. The goal of the task was to automatically assign ICD10 codes (CIE-10, in Spanish) to clinical case documents, being evaluated against manually generated ICD10 codifications. The CodiEsp corpus was selected manually by practicing physicians and clinical documentalists and annotated by clinical coding professionals meeting strict quality criteria. They reached an inter-annotator agreement of 88.6% for diagnosis coding, 88.9% for procedure coding and 80.5% for the textual reference annotation.The final collection of 1,000 clinical cases that make up the corpus had a total of 16,504 sentences and 396,988 words. All documents are in Spanish language and CIE10 is the coding terminology (the Spanish version of ICD10-CM and ICD10-PCS). The CodiEsp corpus has been randomly sampled into three subsets. The train set contains 500 clinical cases, while the development and test sets have 250 clinical cases each. In addition to these, a collection of 176,294 abstracts from Lilacs and Ibecs with the corresponding ICD10 codes (ICD10-CM and ICD10-PCS) was provided by the task organizers. Every abstract has at least one associated code, with an average of 2.5 ICD10 codes per abstract.

The CodiEsp track was divided into three sub-tracks (2 main and 1 exploratory):

CodiEsp-D: The Diagnosis Coding sub-task, which requires automatic ICD10-CM [CIE10-Diagnóstico] code assignment.

CodiEsp-P: The Procedure Coding sub-task, which requires automatic ICD10-PCS [CIE10-Procedimiento] code assignment.

CodiEsp-X: The Explainable AI exploratory sub-task, which requires to submit the reference to the predicted codes (both ICD10-CM and ICD10-PCS). The goal of this novel task was not only to predict the correct codes but also to present the reference in the text that supports the code predictions.

For further information, please visit https://temu.bsc.es/codiesp or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/codiesp/`

URL: `https://temu.bsc.es/codiesp/`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `KB TEXT`

Splits: train

CodiEsp (Procedure Coding) Data Card

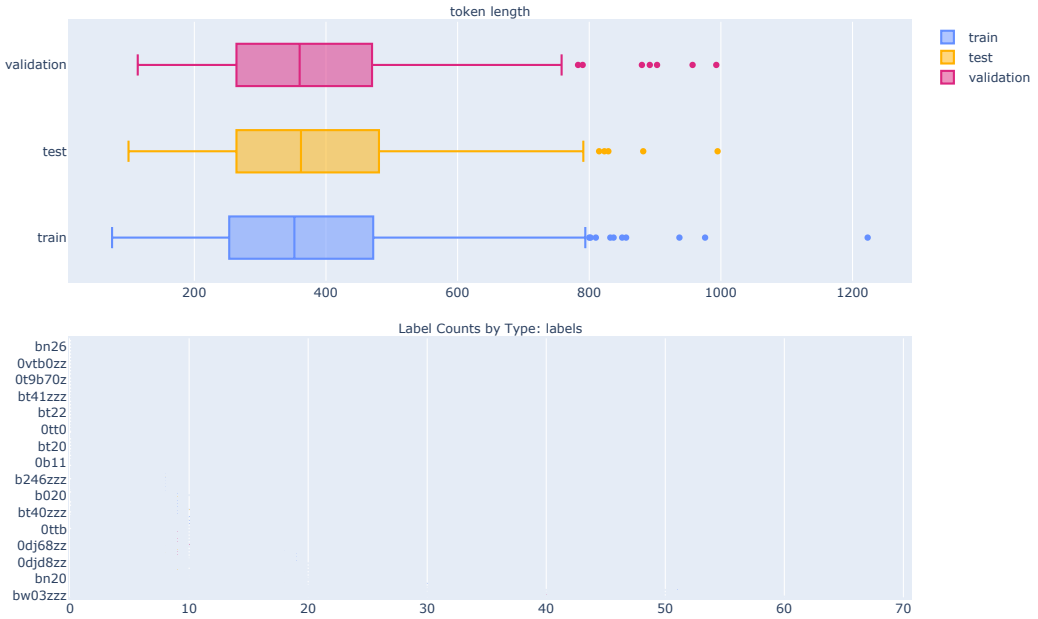

Figure 53: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Synthetic corpus of 1,000 manually selected clinical case studies in Spanish that was designed for the Clinical Case Coding in Spanish Shared Task, as part of the CLEF 2020 conference. The goal of the task was to automatically assign ICD10 codes (CIE-10, in Spanish) to clinical case documents, being evaluated against manually generated ICD10 codifications. The CodiEsp corpus was selected manually by practicing physicians and clinical documentalists and annotated by clinical coding professionals meeting strict quality criteria. They reached an inter-annotator agreement of 88.6% for diagnosis coding, 88.9% for procedure coding and 80.5% for the textual reference annotation.The final collection of 1,000 clinical cases that make up the corpus had a total of 16,504 sentences and 396,988 words. All documents are in Spanish language and CIE10 is the coding terminology (the Spanish version of ICD10-CM and ICD10-PCS). The CodiEsp corpus has been randomly sampled into three subsets. The train set contains 500 clinical cases, while the development and test sets have 250 clinical cases each. In addition to these, a collection of 176,294 abstracts from Lilacs and Ibecs with the corresponding ICD10 codes (ICD10-CM and ICD10-PCS) was provided by the task organizers. Every abstract has at least one associated code, with an average of 2.5 ICD10 codes per abstract.

The CodiEsp track was divided into three sub-tracks (2 main and 1 exploratory):

CodiEsp-D: The Diagnosis Coding sub-task, which requires automatic ICD10-CM [CIE10-Diagnóstico] code assignment.

CodiEsp-P: The Procedure Coding sub-task, which requires automatic ICD10-PCS [CIE10-Procedimiento] code assignment.

CodiEsp-X: The Explainable AI exploratory sub-task, which requires to submit the reference to the predicted codes (both ICD10-CM and ICD10-PCS). The goal of this novel task was not only to predict the correct codes but also to present the reference in the text that supports the code predictions.

For further information, please visit https://temu.bsc.es/codiesp or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/codiesp/`

URL: `https://temu.bsc.es/codiesp/`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: text classification

Schemas: `TEXT`

Splits: train, test, validation

CodiEsp (Explainable AI) Data Card

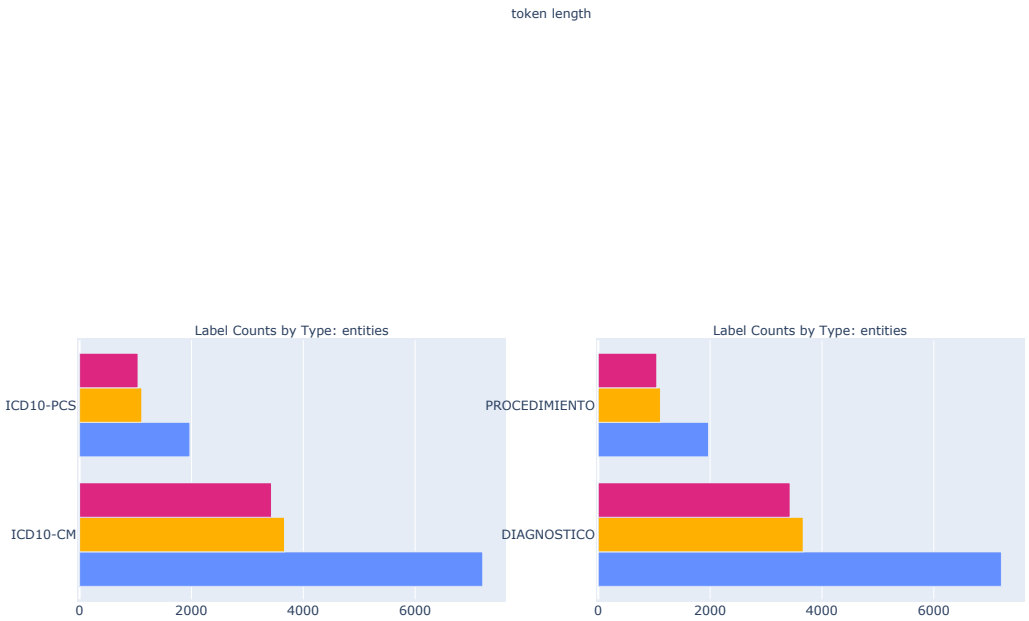

Figure 54: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Synthetic corpus of 1,000 manually selected clinical case studies in Spanish that was designed for the Clinical Case Coding in Spanish Shared Task, as part of the CLEF 2020 conference. The goal of the task was to automatically assign ICD10 codes (CIE-10, in Spanish) to clinical case documents, being evaluated against manually generated ICD10 codifications. The CodiEsp corpus was selected manually by practicing physicians and clinical documentalists and annotated by clinical coding professionals meeting strict quality criteria. They reached an inter-annotator agreement of 88.6% for diagnosis coding, 88.9% for procedure coding and 80.5% for the textual reference annotation.The final collection of 1,000 clinical cases that make up the corpus had a total of 16,504 sentences and 396,988 words. All documents are in Spanish language and CIE10 is the coding terminology (the Spanish version of ICD10-CM and ICD10-PCS). The CodiEsp corpus has been randomly sampled into three subsets. The train set contains 500 clinical cases, while the development and test sets have 250 clinical cases each. In addition to these, a collection of 176,294 abstracts from Lilacs and Ibecs with the corresponding ICD10 codes (ICD10-CM and ICD10-PCS) was provided by the task organizers. Every abstract has at least one associated code, with an average of 2.5 ICD10 codes per abstract.

The CodiEsp track was divided into three sub-tracks (2 main and 1 exploratory):

CodiEsp-D: The Diagnosis Coding sub-task, which requires automatic ICD10-CM [CIE10-Diagnóstico] code assignment.

CodiEsp-P: The Procedure Coding sub-task, which requires automatic ICD10-PCS [CIE10-Procedimiento] code assignment.

CodiEsp-X: The Explainable AI exploratory sub-task, which requires to submit the reference to the predicted codes (both ICD10-CM and ICD10-PCS). The goal of this novel task was not only to predict the correct codes but also to present the reference in the text that supports the code predictions.

For further information, please visit https://temu.bsc.es/codiesp or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/codiesp/`

URL: `https://temu.bsc.es/codiesp/`

Licensing: Creative Commons Attribution 4.0 International

Languages: Spanish

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB TEXT`

Splits: train, test, validation

CORD-NER Data Card

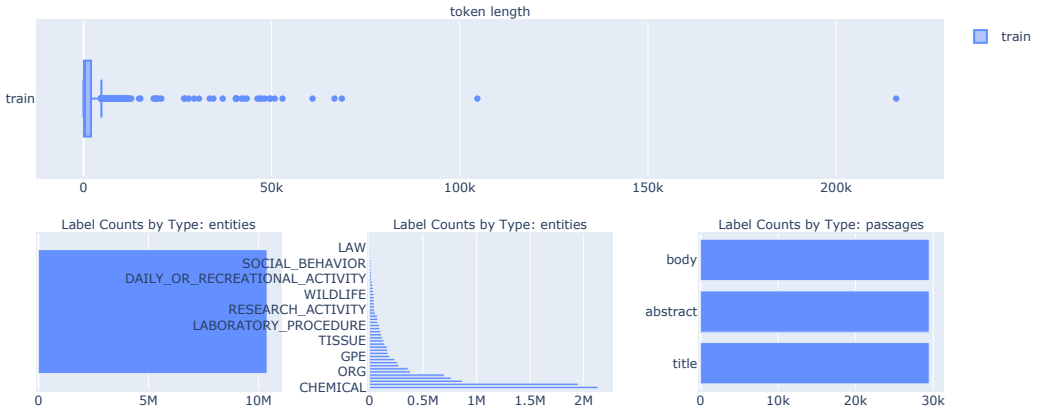

Figure 55: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: CORD-NER dataset covers 75 fine-grained entity types: In addition to the common biomedical entity types (e.g., genes, chemicals and diseases), it covers many new entity types related explicitly to the COVID-19 studies (e.g., coronaviruses, viral proteins, evolution, materials, substrates and immune responses), which may benefit research on COVID-19 related virus, spreading mechanisms, and potential vaccines. CORD-NER annotation is a combination of four sources with different NER methods.

Homepage: `https://xuanwang91.github.io/2020-03-20-cord19-ner/`

URL: `https://xuanwang91.github.io/2020-03-20-cord19-ner/`

Licensing: Custom license

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

CT-EBM-SP (Abstracts) Data Card

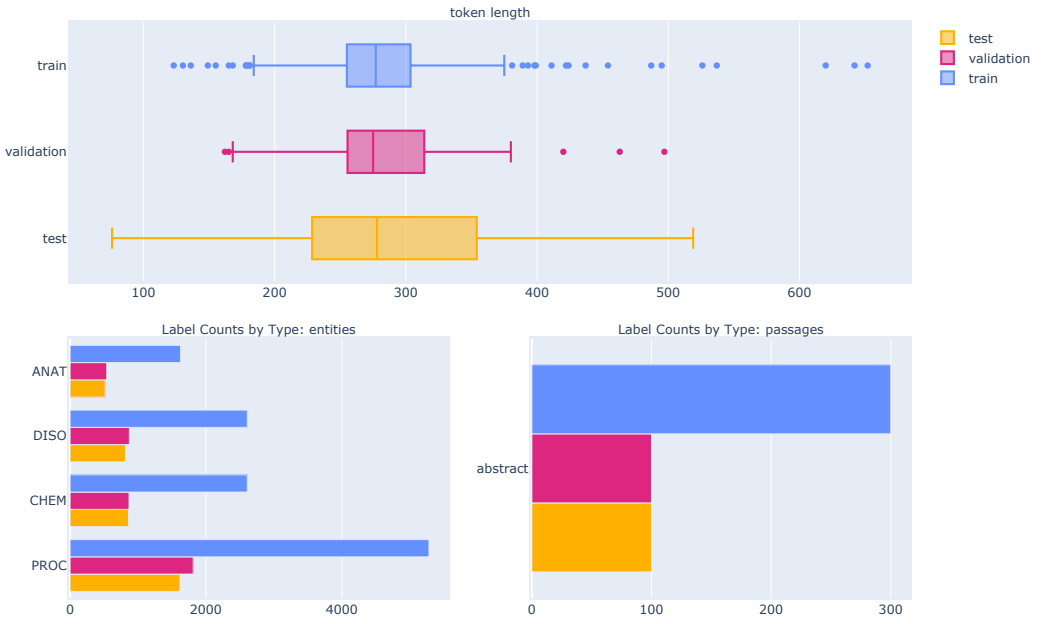

Figure 56: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The "abstracts" subset of the Clinical Trials for Evidence-Based Medicine in Spanish(CT-EBM-SP) corpus contains 500 abstracts of clinical trial studies in Spanish,published in journals with a Creative Commons license. Most were downloaded from the SciELO repository and free abstracts in PubMed. Abstracts were retrieved with the query:Clinical Trial[ptyp] AND “loattrfree full text”[sb] AND “spanish”[la].(Information collected from 10.1186/s12911-021-01395-z)

Homepage: `http://www.lllf.uam.es/ESP/nlpmedterm_en.html`

URL: `http://www.lllf.uam.es/ESP/nlpmedterm_en.html`

Licensing: CC_BY_NC_4p0

Languages: Spanish

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

CT-EBM-SP (Eudract) Data Card

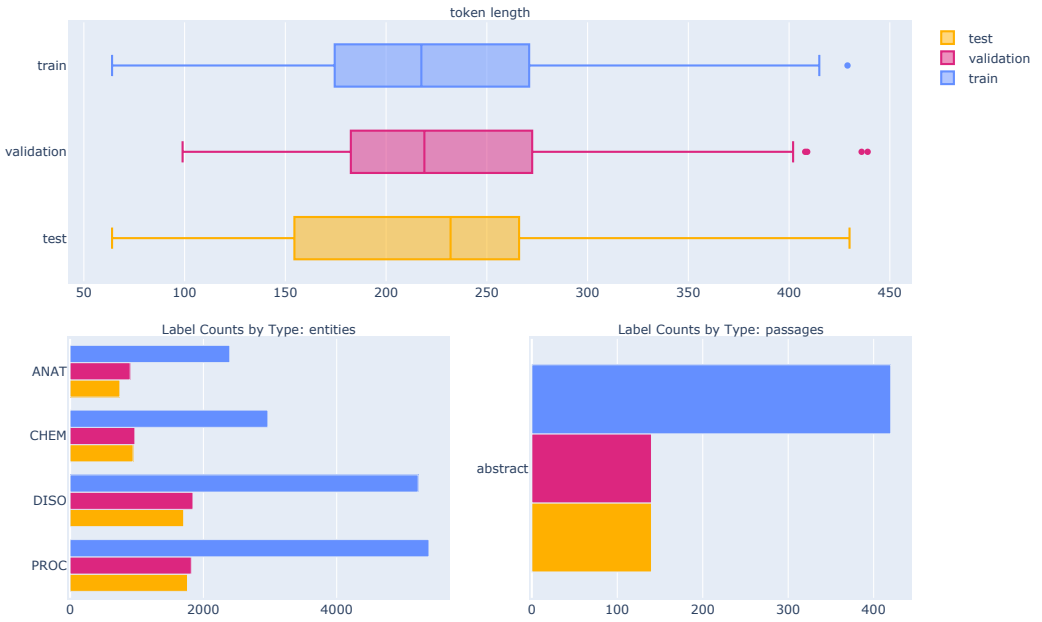

Figure 57: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The "abstracts" subset of the Clinical Trials for Evidence-Based Medicine in Spanish(CT-EBM-SP) corpus contains 500 abstracts of clinical trial studies in Spanish,published in journals with a Creative Commons license. Most were downloaded from the SciELO repository and free abstracts in PubMed.Abstracts were retrieved with the query:Clinical Trial[ptyp] AND “loattrfree full text”[sb] AND “spanish”[la].(Information collected from 10.1186/s12911-021-01395-z)

Homepage: `http://www.lllf.uam.es/ESP/nlpmedterm_en.html`

URL: `http://www.lllf.uam.es/ESP/nlpmedterm_en.html`

Licensing: CC_BY_NC_4p0

Languages: Spanish

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

DDI corpus Data Card

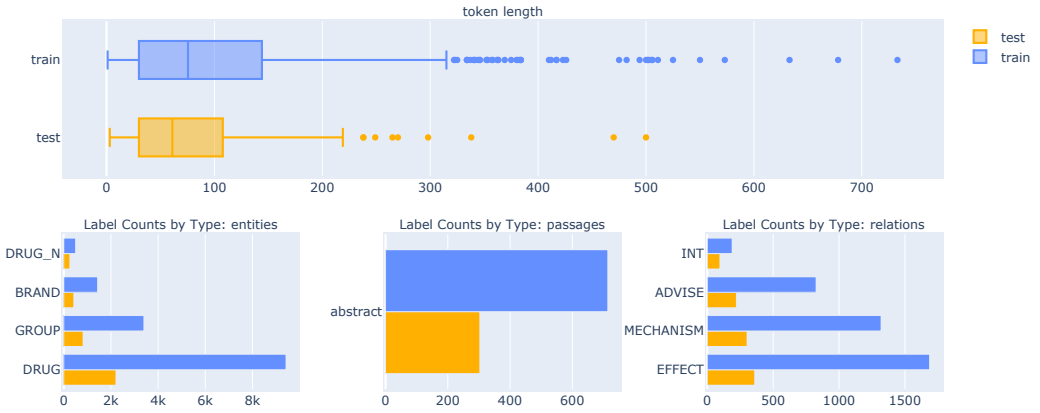

Figure 58: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The DDI corpus has been manually annotated with drugs and pharmacokinetics and pharmacodynamics interactions. It contains 1025 documents from two different sources: DrugBank database and MedLine.

Homepage: `https://github.com/isegura/DDICorpus`

URL: `https://github.com/isegura/DDICorpus`

Licensing: CC_BY_NC_4p0

Languages: English

Tasks: named entity recognition, relation extraction

Schemas: `KB`

Splits: train, test

DIANN Data Card

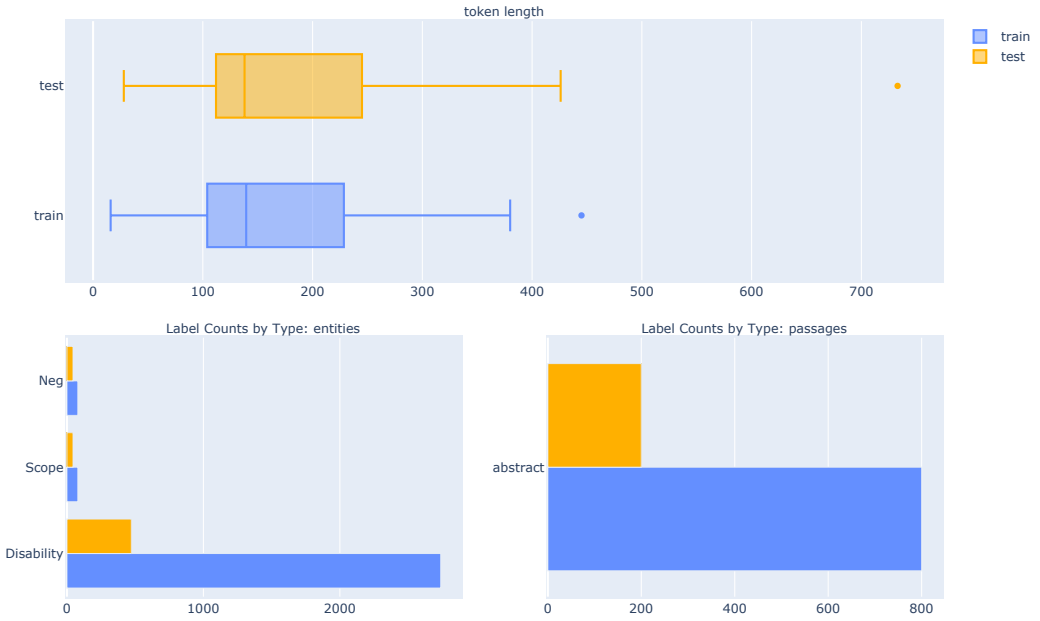

Figure 59: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description DIANN's corpus consists of a collection of 500 abstracts from Elsevier journal papers related to the biomedical domain, where both Spanish and English versions are available.This dataset contains annotations for disabilities appearing in these abstracts,usually expressed either with a specific word, such as "blindness", or as the limitation or lack of a human function, such as "lack of vision".

(Spanish) El corpus DIANN se compone de una colección de 500 resúmenes de artículos de revista Elsevierdel ámbito biomédico, con versiones en español e inglés.Este conjunto de datos contiene anotaciones para discapacidades que aparecen en dichos resúmenes,expresadas por medio de palabras específicas, como "ablepsia", or como la limitación o falta deuna función, como "falta de visión".

Homepage: `http://nlp.uned.es/diann/`

URL: `http://nlp.uned.es/diann/`

Licensing: UNKNOWN

Languages: English, Spanish

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test

DIANN Data Card

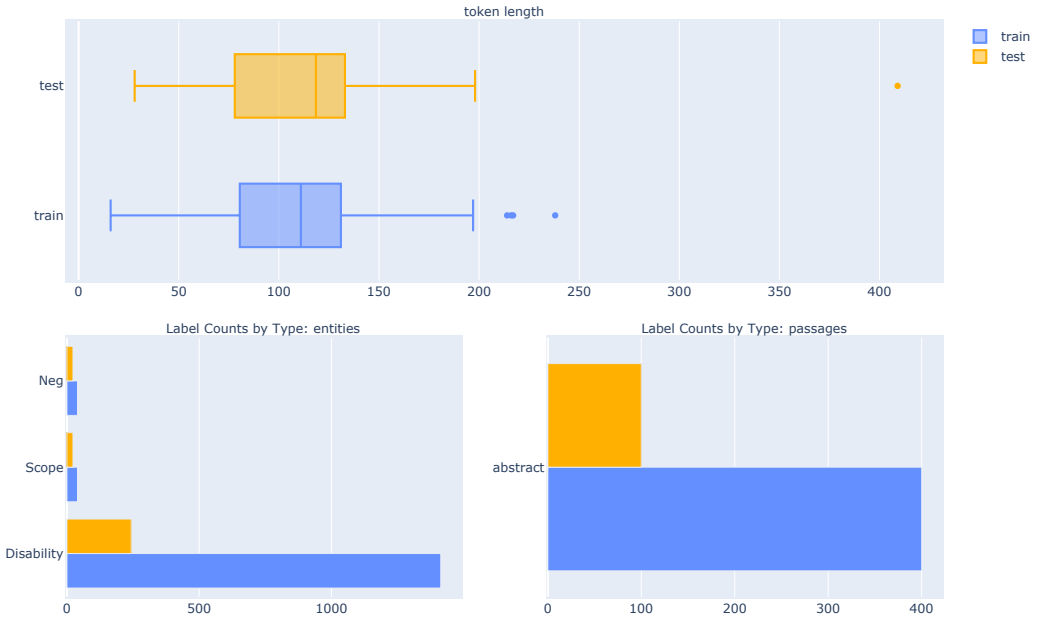

Figure 60: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: DIANN's corpus consists of a collection of 500 abstracts from Elsevier journal papers related to the biomedical domain, where both Spanish and English versions are available.This dataset contains annotations for disabilities appearing in these abstracts,usually expressed either with a specific word, such as "blindness", or as the limitation or lack of a human function, such as "lack of vision".

(Spanish) El corpus DIANN se compone de una colección de 500 resúmenes de artículos de revista Elsevierdel ámbito biomédico, con versiones en español e inglés.Este conjunto de datos contiene anotaciones para discapacidades que aparecen en dichos resúmenes,expresadas por medio de palabras específicas, como "ablepsia", or como la limitación o falta deuna función, como "falta de visión".

Homepage: `http://nlp.uned.es/diann/`

URL: `http://nlp.uned.es/diann/`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test

DIANN Data Card

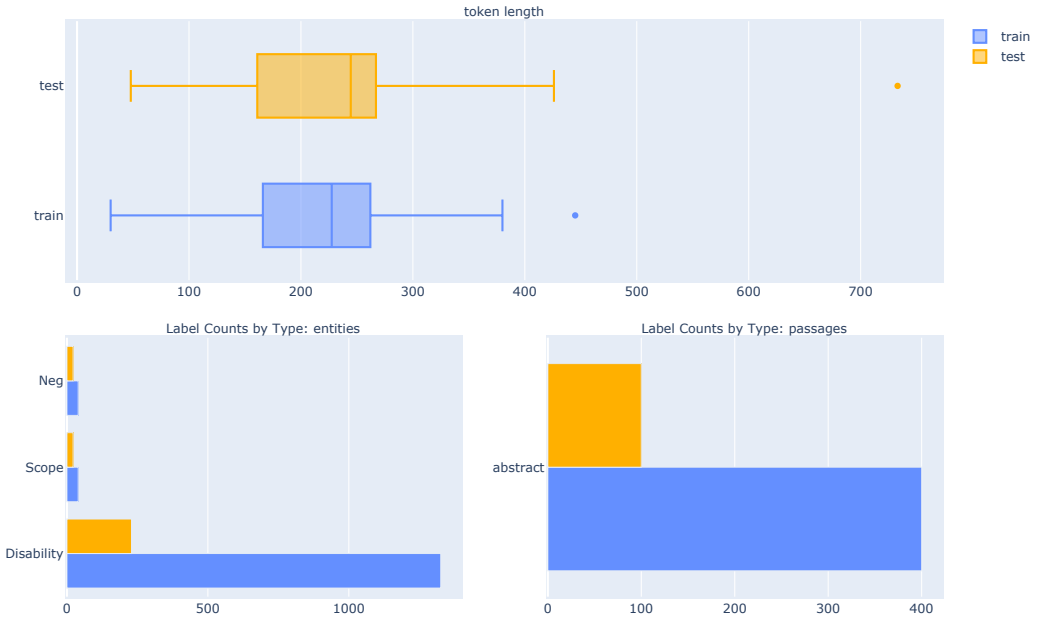

Figure 61: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: DIANN's corpus consists of a collection of 500 abstracts from Elsevier journal papers related to the biomedical domain, where both Spanish and English versions are available.This dataset contains annotations for disabilities appearing in these abstracts,usually expressed either with a specific word, such as "blindness", or as the limitation or lack of a human function, such as "lack of vision".

(Spanish) El corpus DIANN se compone de una colección de 500 resúmenes de artículos de revista Elsevierdel ámbito biomédico, con versiones en español e inglés.Este conjunto de datos contiene anotaciones para discapacidades que aparecen en dichos resúmenes,expresadas por medio de palabras específicas, como "ablepsia", or como la limitación o falta deuna función, como "falta de visión".

Homepage: `http://nlp.uned.es/diann/`

URL: `http://nlp.uned.es/diann/`

Licensing: UNKNOWN

Languages: Spanish

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test

DisTEMIST Data Card

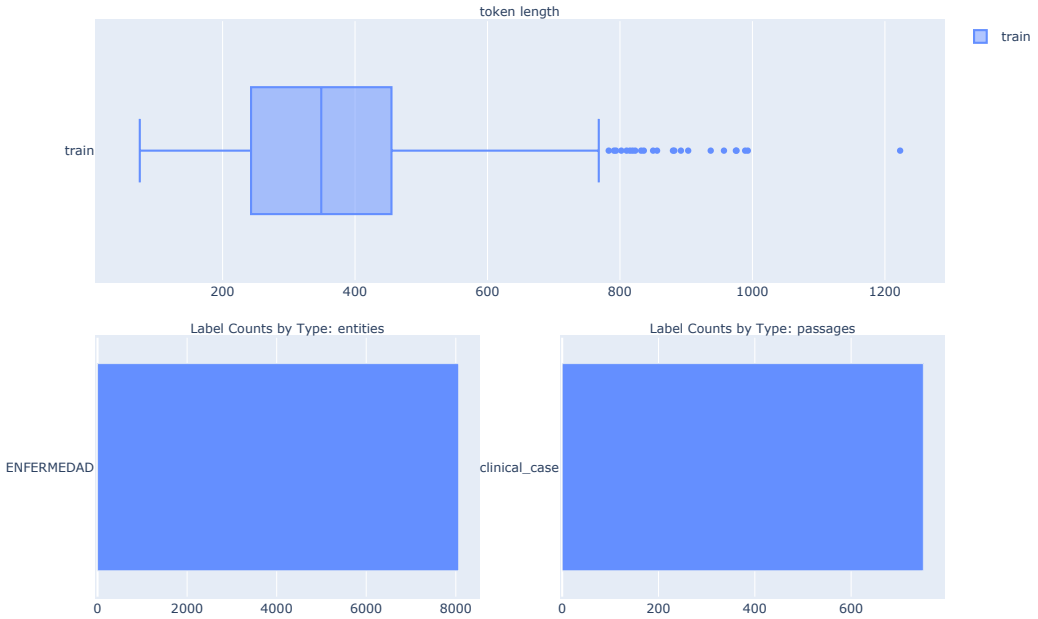

Figure 62: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The DisTEMIST corpus is a collection of 1000 clinical cases with disease annotations linked with Snomed-CT concepts.All documents are released in the context of the BioASQ DisTEMIST track for CLEF 2022.

Homepage: `https://zenodo.org/record/6458455`

URL: `https://zenodo.org/record/6458455`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

EBM NLP Data Card

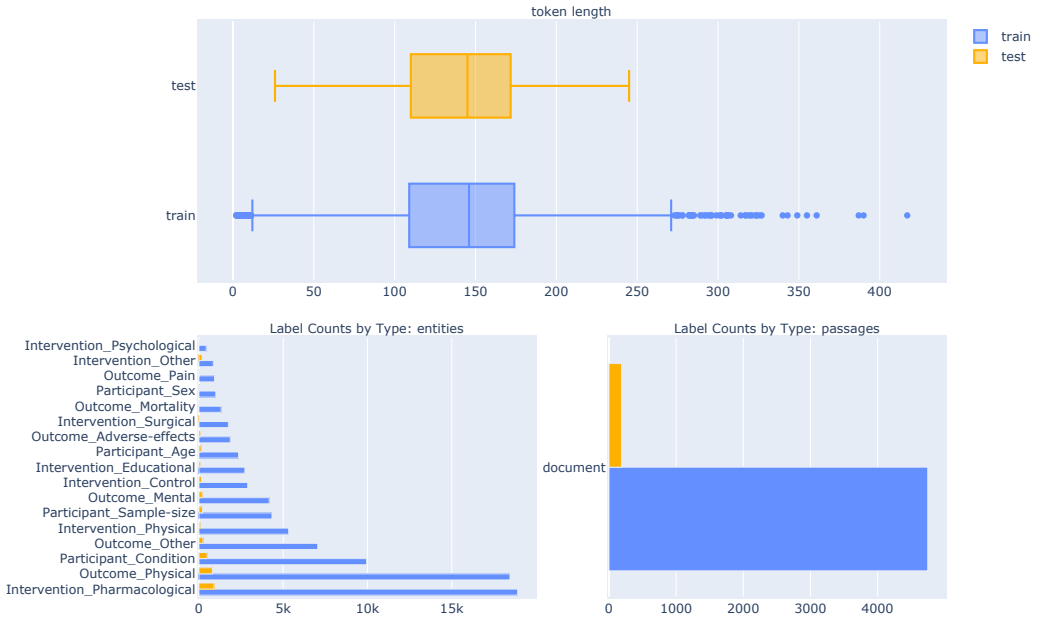

Figure 63: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: This corpus release contains 4,993 abstracts annotated with (P)articipants, (I)nterventions, and (O)utcomes. Training labels are sourced from AMT workers and aggregated to reduce noise. Test labels are collected from medical professionals.

Homepage: `https://github.com/bepnye/EBM-NLP`

URL: `https://github.com/bepnye/EBM-NLP`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test

EHR-Rel Data Card

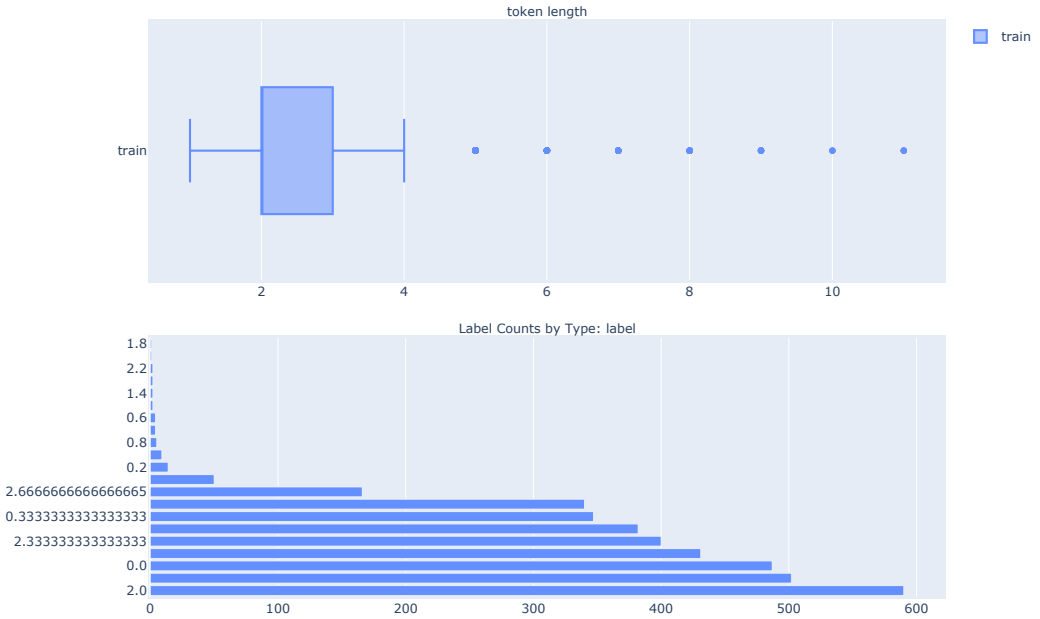

Figure 64: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: EHR-Rel is a novel open-source1 biomedical concept relatedness dataset consisting of 3630 concept pairs, six times more than the largest existing dataset. Instead of manually selecting and pairing concepts as done in previous work, the dataset is sampled from EHRs to ensure concepts are relevant for the EHR concept retrieval task.A detailed analysis of the concepts in the dataset reveals a far larger coverage compared to existing datasets.

Homepage: `https://github.com/babylonhealth/EHR-Rel`

URL: `https://github.com/babylonhealth/EHR-Rel`

Licensing: APACHE_2p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

EU-ADR Data Card

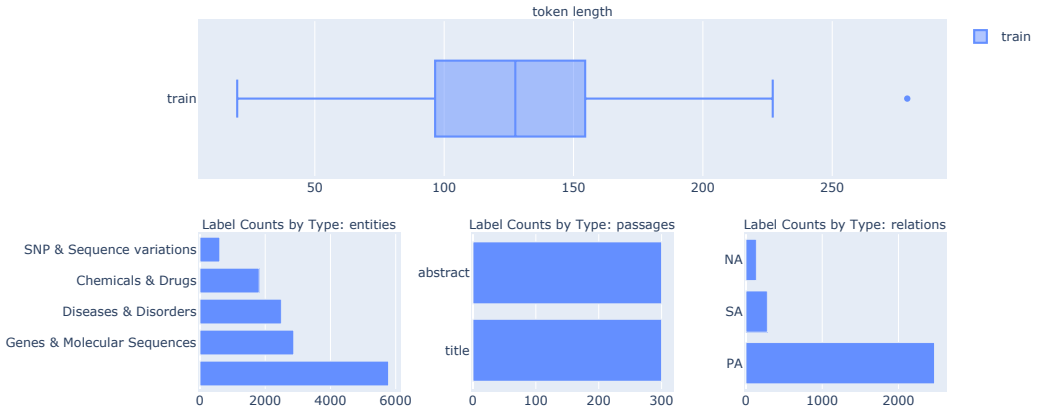

Figure 65: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Corpora with specific entities and relationships annotated are essential to train and evaluate text-mining systems that are developed to extract specific structured information from a large corpus. In this paper we describe an approach where a named-entity recognition system produces a first annotation and annotators revise this annotation using a web-based interface. The agreement figures achieved show that the inter-annotator agreement is much better than the agreement with the system provided annotations. The corpus has been annotated for drugs, disorders, genes and their inter-relationships. For each of the drug–disorder, drug–target, and target–disorder relations three experts have annotated a set of 100 abstracts. These annotated relationships will be used to train and evaluate text-mining software to capture these relationships in texts.

Homepage: `https://www.sciencedirect.com/science/article/pii/S1532046412000573`

URL: `https://www.sciencedirect.com/science/article/pii/S1532046412000573`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition, relation extraction

Schemas: `KB`

Splits: train

Evidence Inference Data Card

Figure 66: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The dataset consists of biomedical articles describing randomized control trials (RCTs) that compare multiple treatments. Each of these articles will have multiple questions, or 'prompts' associated with them.These prompts will ask about the relationship between an intervention and comparator with respect to an outcome,as reported in the trial. For example, a prompt may ask about the reported effects of aspirin as compared to placebo on the duration of headaches. For the sake of this task, we assume that a particular article will report that the intervention of interest either significantly increased, significantly decreased or had significant effect on the outcome, relative to the comparator.

Homepage: `https://github.com/jayded/evidence-inference`

URL: `https://github.com/jayded/evidence-inference`

Licensing: MIT

Languages: English

Tasks: textual entailment

Schemas: `TE`

Splits: train, validation, test

GAD Data Card

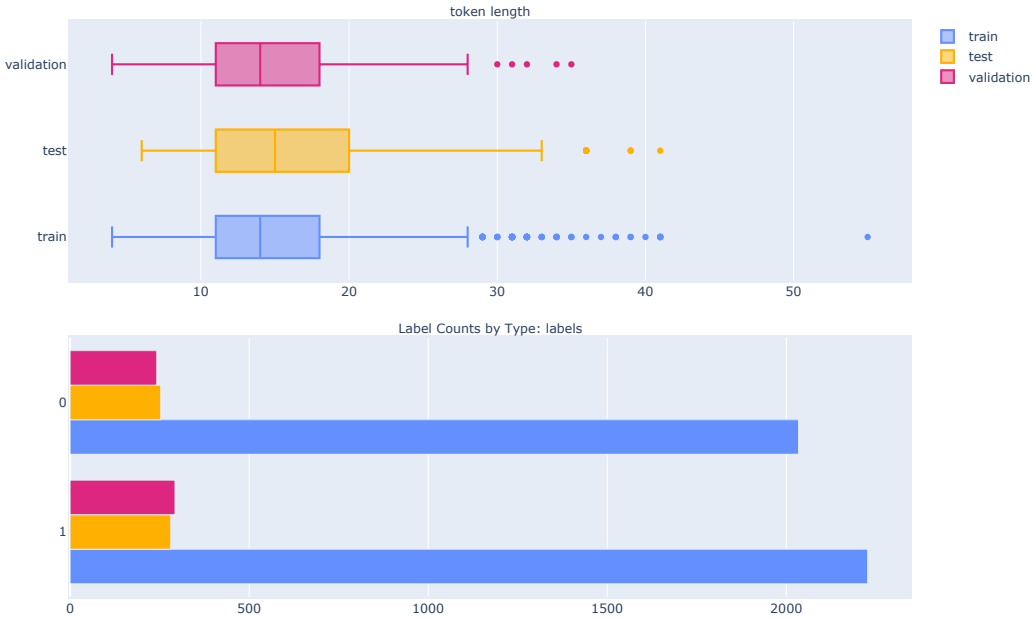

Figure 67: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: A corpus identifying associations between genes and diseases by a semi-automaticannotation procedure based on the Genetic Association Database

Homepage: `https://github.com/dmis-lab/biobert`

URL: `https://github.com/dmis-lab/biobert`

Licensing: CC_BY_4p0

Languages: English

Tasks: text classification

Schemas: `TEXT`

Splits: train, validation, test

GENETAG Data Card

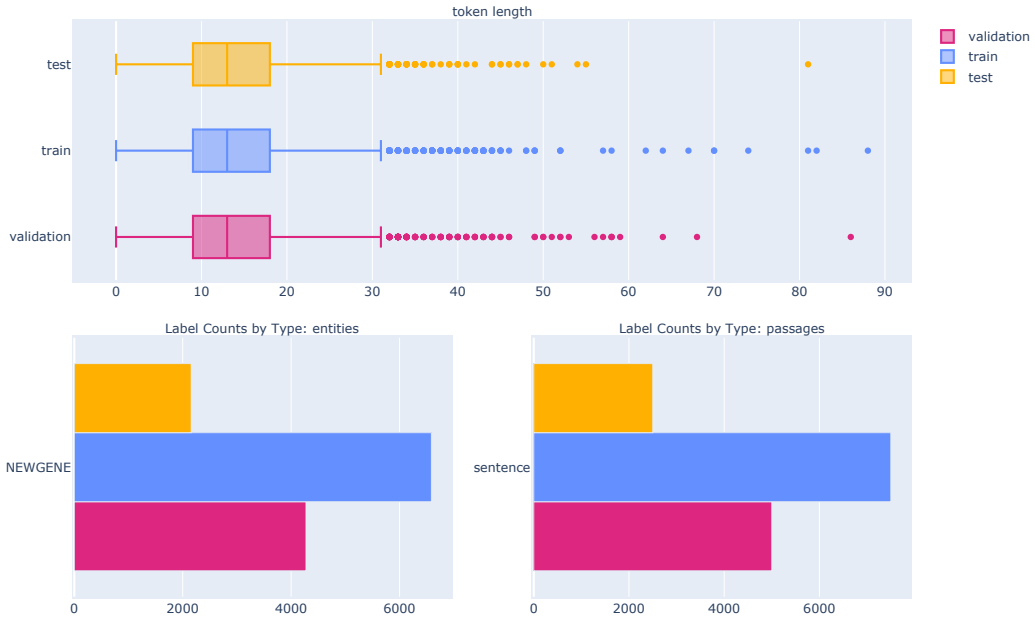

Figure 68: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Named entity recognition (NER) is an important first step for text mining the biomedical literature.Evaluating the performance of biomedical NER systems is impossible without a standardized test corpus.The annotation of such a corpus for gene/protein name NER is a difficult process due to the complexity of gene/protein names. We describe the construction and annotation of GENETAG, a corpus of 20K MEDLINE®sentences for gene/protein NER. 15K GENETAG sentences were used for the BioCreAtIvE Task 1A Competition..

Homepage: `https://github.com/openbiocorpora/genetag`

URL: `https://github.com/openbiocorpora/genetag`

Licensing: NCBI_LICENSE

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

GENETAG Data Card

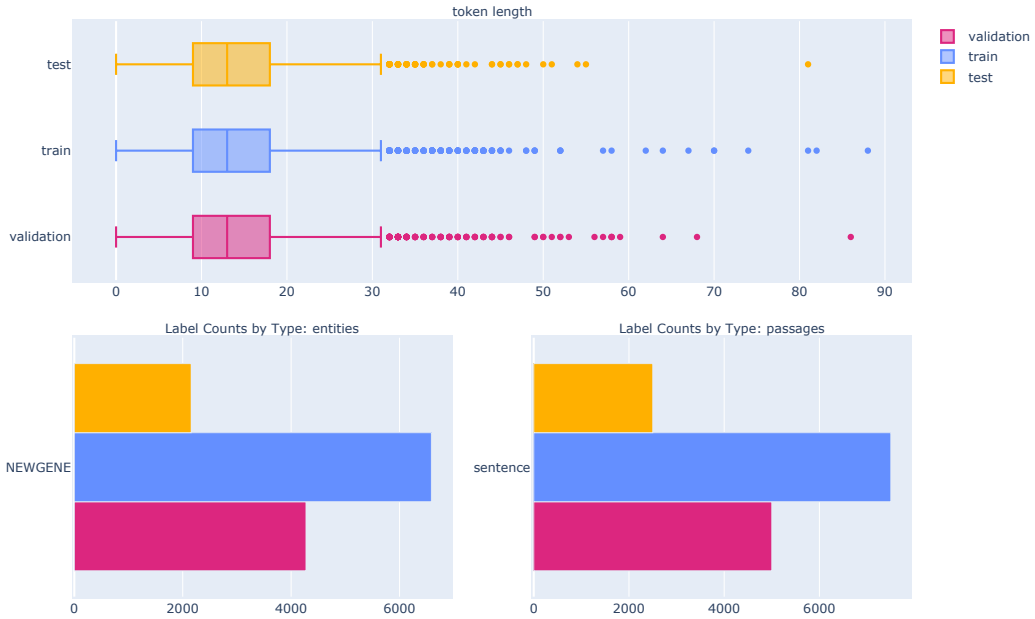

Figure 69: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Named entity recognition (NER) is an important first step for text mining the biomedical literature.Evaluating the performance of biomedical NER systems is impossible without a standardized test corpus.The annotation of such a corpus for gene/protein name NER is a difficult process due to the complexity of gene/protein names. We describe the construction and annotation of GENETAG, a corpus of 20K MEDLINE®sentences for gene/protein NER. 15K GENETAG sentences were used for the BioCreAtIvE Task 1A Competition..

Homepage: `https://github.com/openbiocorpora/genetag`

URL: `https://github.com/openbiocorpora/genetag`

Licensing: NCBI_LICENSE

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

PTM Events corpus Data Card

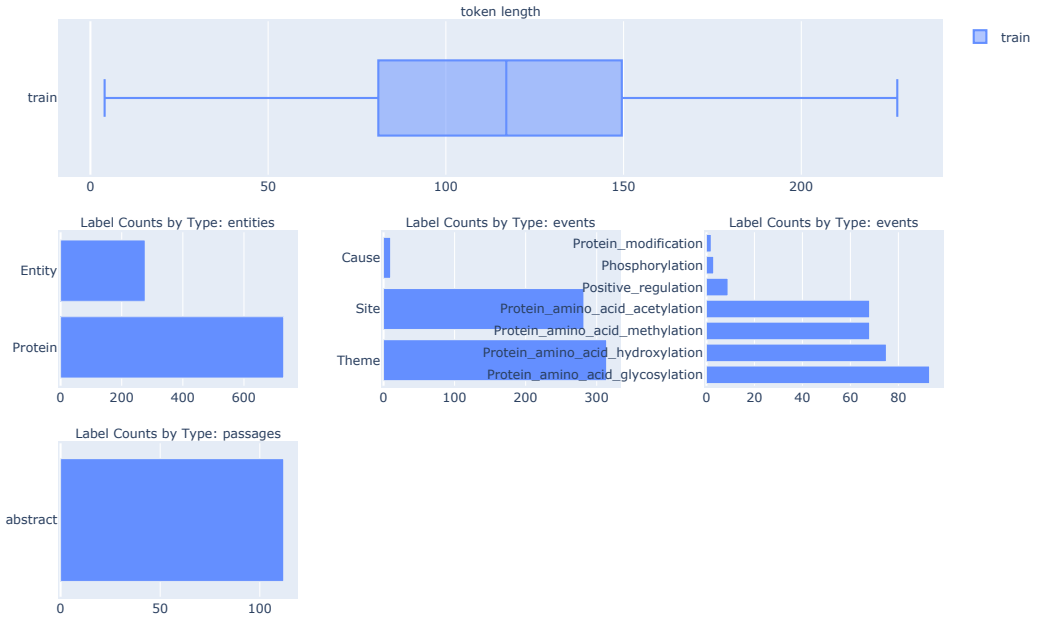

Figure 70: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Post-translational-modifications (PTM), amino acid modifications of proteins after translation, are one of the posterior processes of protein biosynthesis for many proteins, and they are critical for determining protein function such as its activity state, localization, turnover and interactions with other biomolecules. While there have been many studies of information extraction targeting individual PTM types, there was until recently little effort to address extraction of multiple PTM types at once in a unified framework.

Homepage: `http://www.geniaproject.org/other-corpora/ptm-event-corpus`

URL: `http://www.geniaproject.org/other-corpora/ptm-event-corpus`

Licensing: GENIA_PROJECT_LICENSE

Languages: English

Tasks: named entity recognition, event extraction, coreference resolution

Schemas: `KB`

Splits: train

GENIA Relation Corpus Data Card

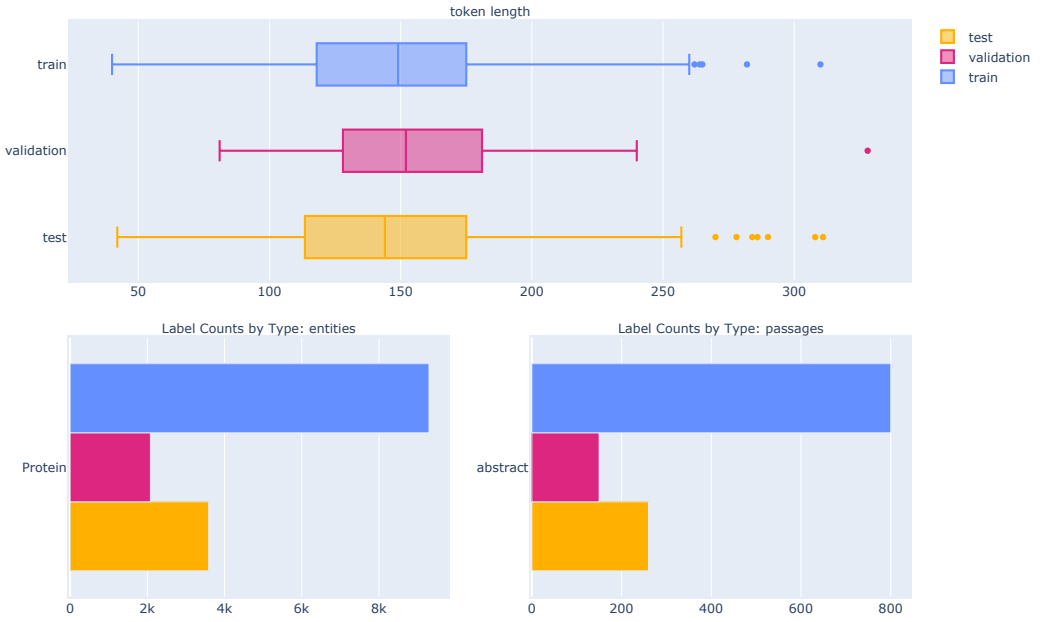

Figure 71: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The extraction of various relations stated to hold between biomolecular entities is one of the most frequently addressed information extraction tasks in domain studies. Typical relation extraction targets involve protein-protein interactions or gene regulatory relations. However, in the GENIA corpus, such associations involving change in the state or properties of biomolecules are captured in the event annotation.The GENIA corpus relation annotation aims to complement the event annotation of the corpus by capturing (primarily)static relations, relations such as part-of that hold between entities without (necessarily) involving change.

Homepage: `http://www.geniaproject.org/genia-corpus/relation-corpus`

URL: `http://www.geniaproject.org/genia-corpus/relation-corpus`

Licensing: GENIA_PROJECT_LICENSE

Languages: English

Tasks: relation extraction

Schemas: `KB`

Splits: train, validation, test

GENIA Term Corpus Data Card

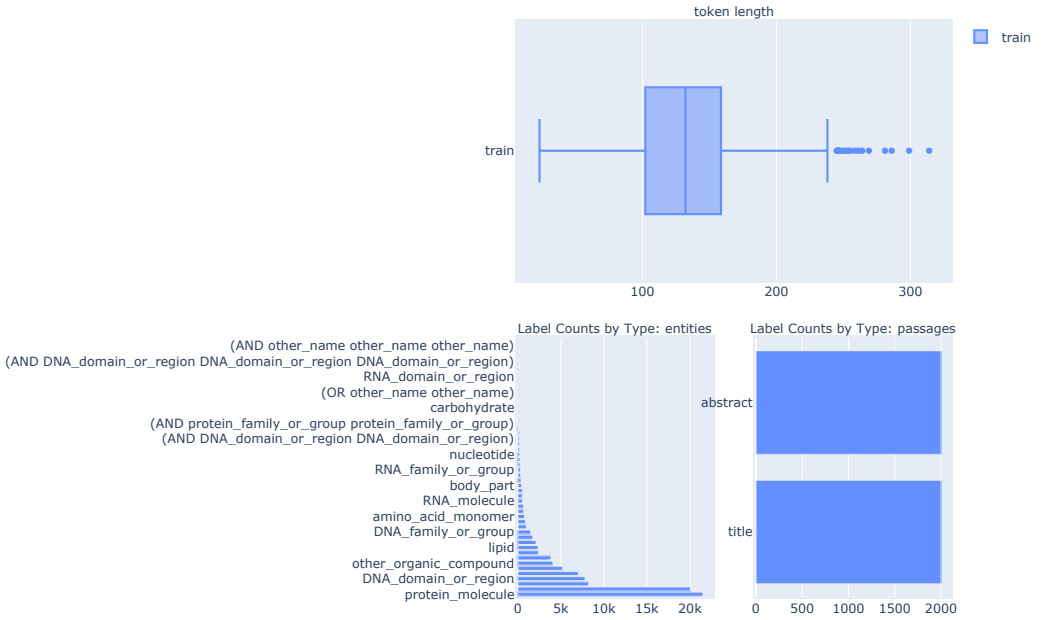

Figure 72: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The identification of linguistic expressions referring to entities of interest in molecular biology such as proteins,genes and cells is a fundamental task in biomolecular text mining. The GENIA technical term annotation covers the identification of physical biological entities as well as other important terms. The corpus annotation covers the full1,999 abstracts of the primary GENIA corpus.

Homepage: `http://www.geniaproject.org/genia-corpus/term-corpus`

URL: `http://www.geniaproject.org/genia-corpus/term-corpus`

Licensing: GENIA_PROJECT_LICENSE

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

GEOkhoj v1 Data Card

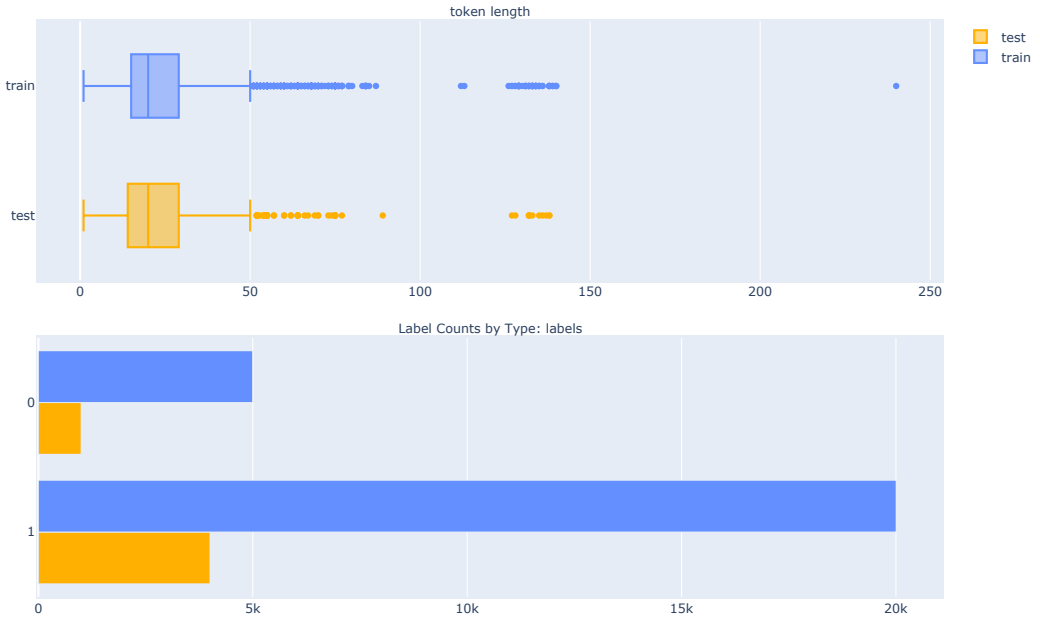

Figure 73: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: GEOKhoj v1 is a annotated corpus of control/perturbation labels for 30,000 samples from Microarray, Transcriptomics and Single cell experiments which are available on the GEO (Gene Expression Omnibus) database

Homepage: `https://github.com/ElucidataInc/GEOKhoj-datasets/tree/main/geokhoj_v1`

URL: `https://github.com/ElucidataInc/GEOKhoj-datasets/tree/main/geokhoj_v1`

Licensing: CC_BY_NC_4p0

Languages: English

Tasks: text classification

Schemas: `TEXT`

Splits: train, test

GNormPlus Data Card

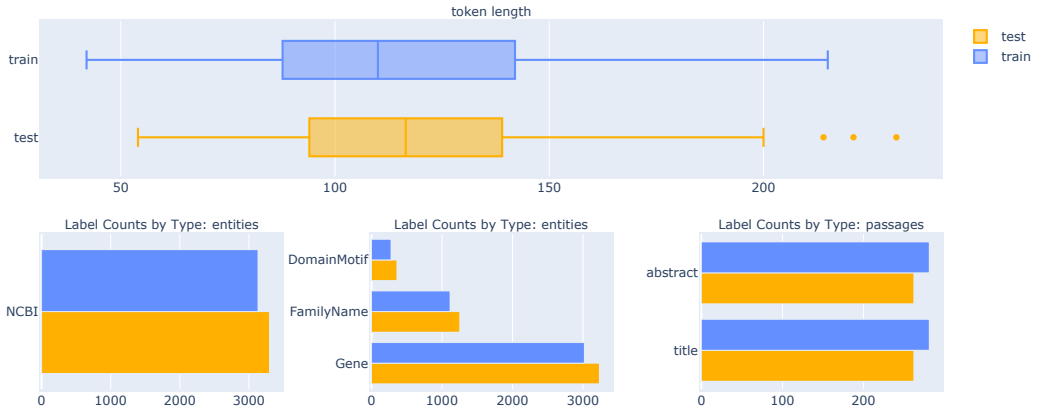

Figure 74: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We re-annotated two existing gene corpora. The BioCreative II GN corpus is a widely used data set for benchmarking GNtools and includes document-level annotations for a total of 543 articles (281 in its training set; and 262 in test).The Citation GIA Test Collection was recently created for gene indexing at the NLM and includes 151 PubMed abstracts with both mention-level and document-level annotations. They are selected because both have a focus on human genes. For both corpora, we added annotations of gene families and protein domains. For the BioCreative GN corpus, we also added mention-level gene annotations. As a result, in our new corpus, there are a total of 694 PubMed articles.PubTator was used as our annotation tool along with BioC formats.

Homepage: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/gnormplus/`

URL: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/gnormplus/`

Licensing: UNKNOWN

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train, test

Hallmarks of Cancer Data Card

Figure 75: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The Hallmarks of Cancer (HOC) Corpus consists of 1852 PubMed publication abstracts manually annotated by experts according to a taxonomy. The taxonomy consists of 37 classes in a hierarchy. Zero or more class labels are assigned to each sentence in the corpus. The labels are found under the "labels"directory, while the tokenized text can be found under "text" directory.The filenames are the corresponding PubMed IDs (PMID).

Homepage: `https://github.com/sb895/Hallmarks-of-Cancer`

URL: `https://github.com/sb895/Hallmarks-of-Cancer`

Licensing: GPL_3p0

Languages: English

Tasks: text classification

Schemas: `TEXT`

Splits: train, test, validation

HPRD50 Data Card

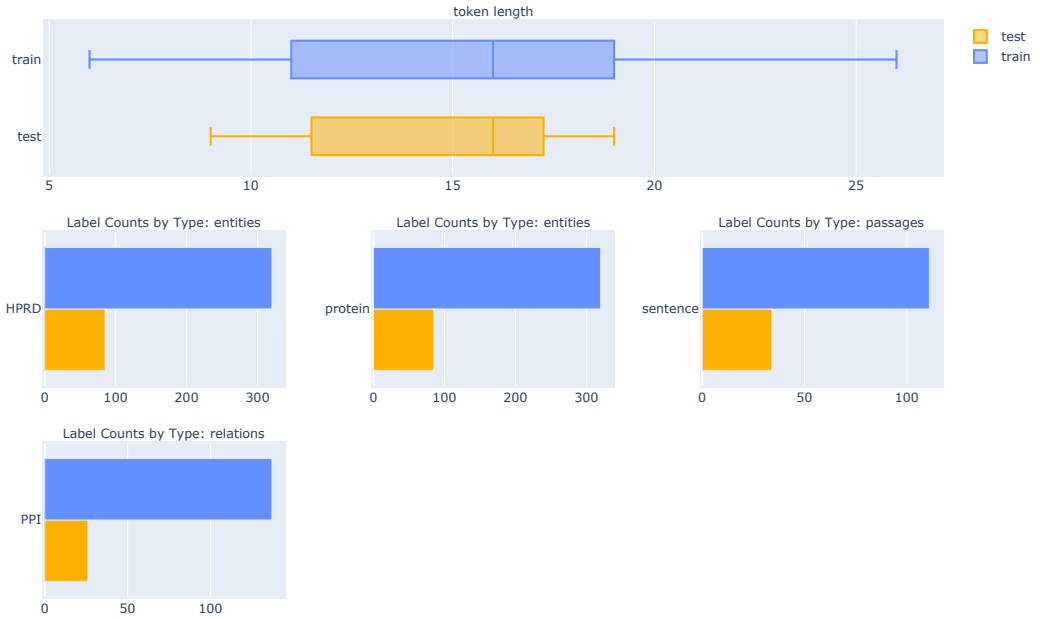

Figure 76: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: HPRD50 is a dataset of randomly selected, hand-annotated abstracts of biomedical papers referenced by the Human Protein Reference Database (HPRD). It is parsed in XML format,splitting each abstract into sentences, and in each sentence there may be entities and interactions between those entities. In this particular dataset, entities are all proteins and interactions are thus protein-protein interactions. Moreover, all entities are normalized to the HPRD database. These normalized terms are stored in each entity's 'type' attribute in the source XML. This means the dataset can determine e.g. that "Janus kinase 2" and "Jak2" are referencing the same normalized entity.Because the dataset contains entities and relations, it is suitable for Named Entity Recognition and Relation Extraction.

Homepage: UNKNOWN

URL: UNKNOWN

Licensing: UNKNOWN

Languages: English

Tasks: relation extraction, named entity recognition

Schemas: `KB`

Splits: train, test

IEPA Data Card

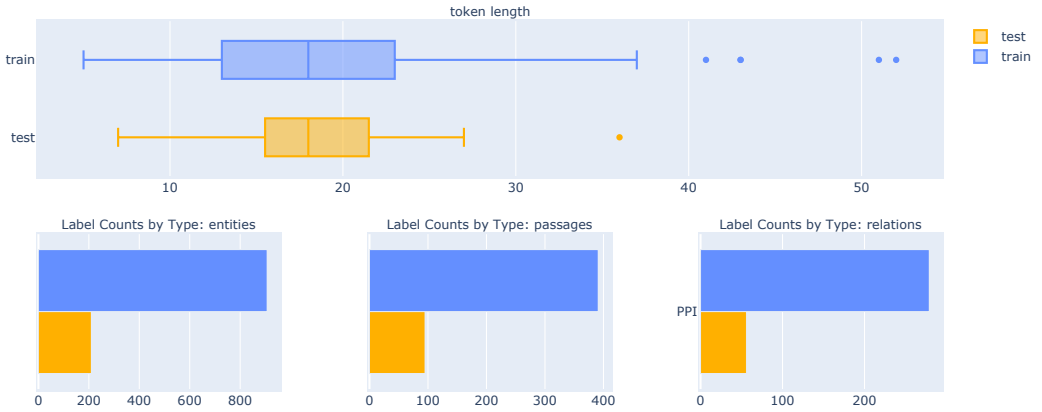

Figure 77: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The IEPA benchmark PPI corpus is designed for relation extraction. It was created from 303 PubMed abstracts,each of which contains a specific pair of co-occurring chemicals.

Homepage: `http://psb.stanford.edu/psb-online/proceedings/psb02/abstracts/p326.html`

URL: `http://psb.stanford.edu/psb-online/proceedings/psb02/abstracts/p326.html`

Licensing: UNKNOWN

Languages: English

Tasks: relation extraction

Schemas: `KB`

Splits: train, test

JNLPBA Data Card

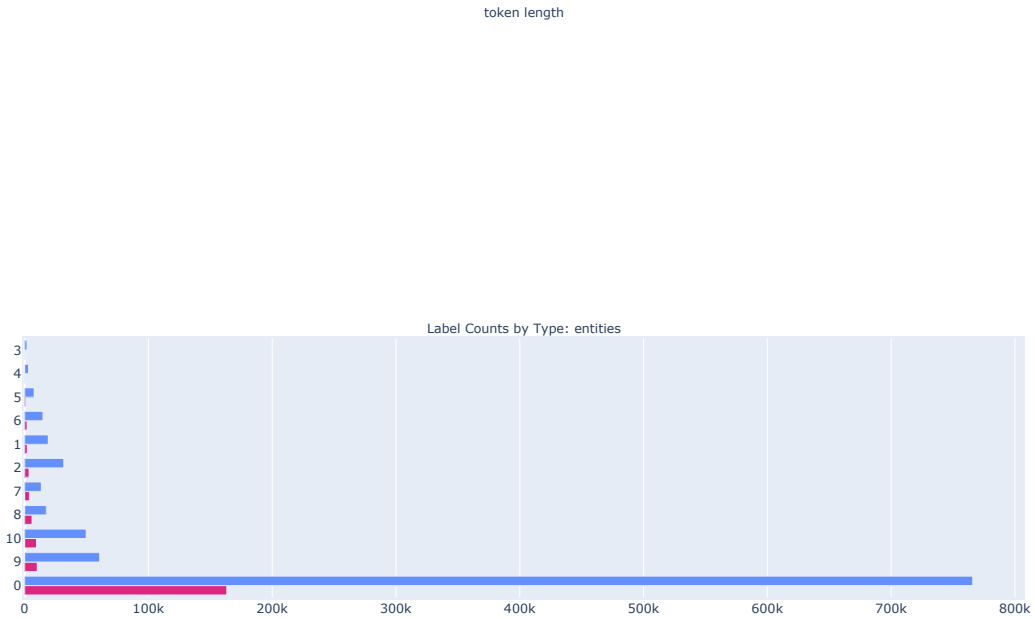

Figure 78: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: NER For Bio-Entities

Homepage: `http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004`

URL: `http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004`

Licensing: CC_BY_3p0

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, validation

LINNAEUS Data Card

Figure 79: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Linnaeus is a novel corpus of full-text documents manually annotated for species mentions.

Homepage: `http://linnaeus.sourceforge.net/`

URL: `http://linnaeus.sourceforge.net/`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

LINNAEUS (Filtered) Data Card

Figure 80: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Linnaeus is a novel corpus of full-text documents manually annotated for species mentions.

Homepage: `http://linnaeus.sourceforge.net/`

URL: `http://linnaeus.sourceforge.net/`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

LLL05 Data Card

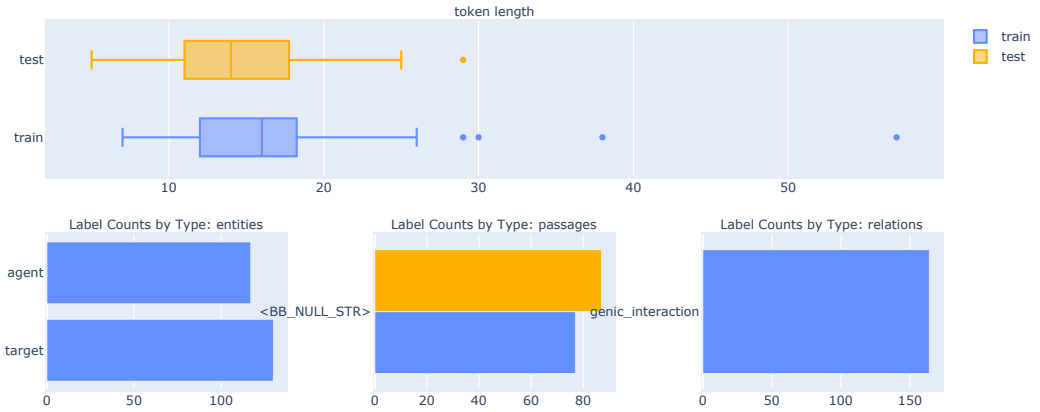

Figure 81: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The LLL05 challenge task is to learn rules to extract protein/gene interactions from biology abstracts from the Medline bibliography database. The goal of the challenge is to test the ability of the participating IE systems to identify the interactions and the gene/proteins that interact. The participants will test their IE patterns on a test set with the aim of extracting the correct agent and target.The challenge focuses on information extraction of gene interactions in Bacillus subtilis. Extracting gene interaction is the most popular event IE task in biology. Bacillus subtilis (Bs) isa model bacterium and many papers have been published on direct gene interactions involved in sporulation. The gene interactions are generally mentioned in the abstract and the full text of the paper is not needed. Extracting gene interaction means, extracting the agent (proteins) and the target (genes) of all couples of genic interactions from sentences.

Homepage: `http://genome.jouy.inra.fr/texte/LLLchallenge`

URL: `http://genome.jouy.inra.fr/texte/LLLchallenge`

Licensing: UNKNOWN

Languages: English

Tasks: relation extraction

Schemas: `KB`

Splits: train, test

Mantra GSC (EMEA German) Data Card

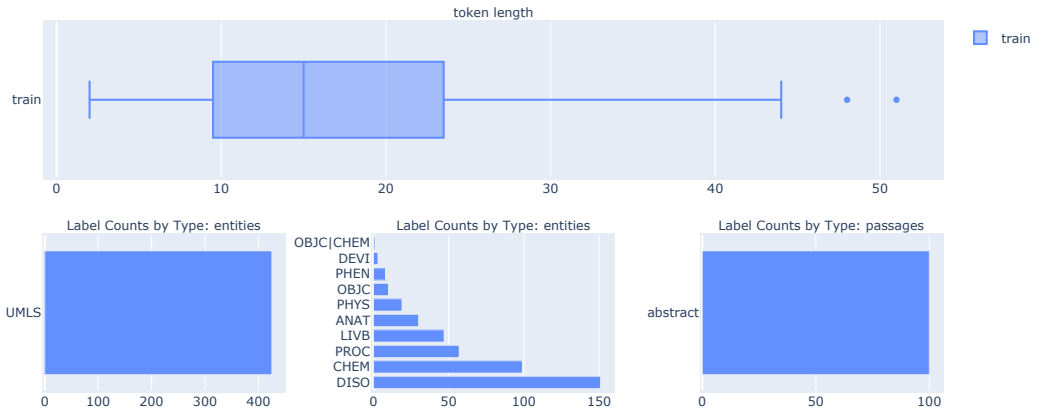

Figure 82: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups.

Homepage: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: German

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (Medline German) Data Card

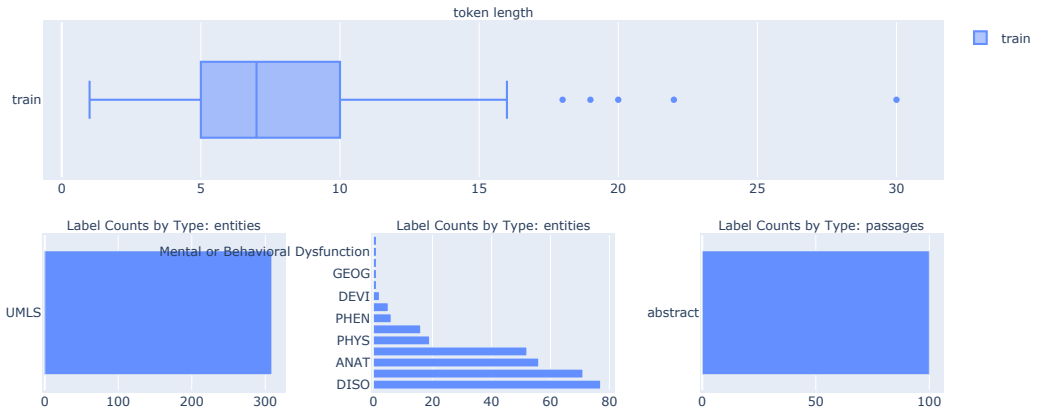

Figure 83: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: German

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (Patents German) Data Card

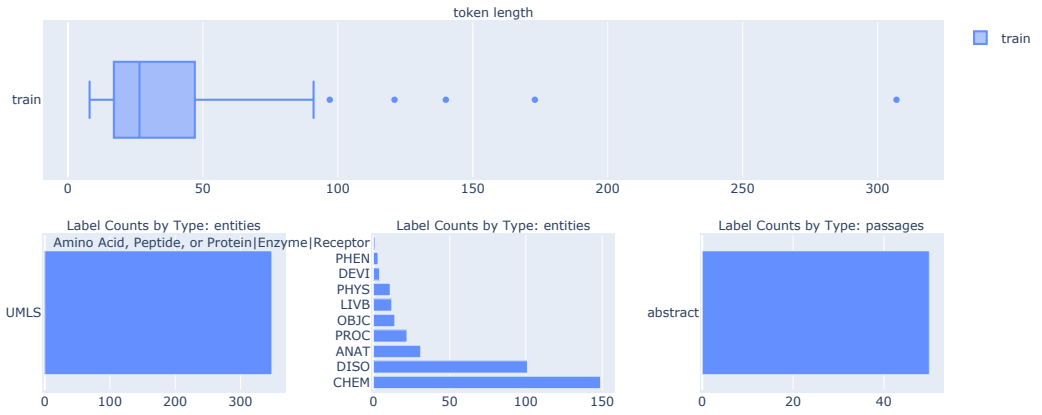

Figure 84: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: German

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (EMEA English) Data Card

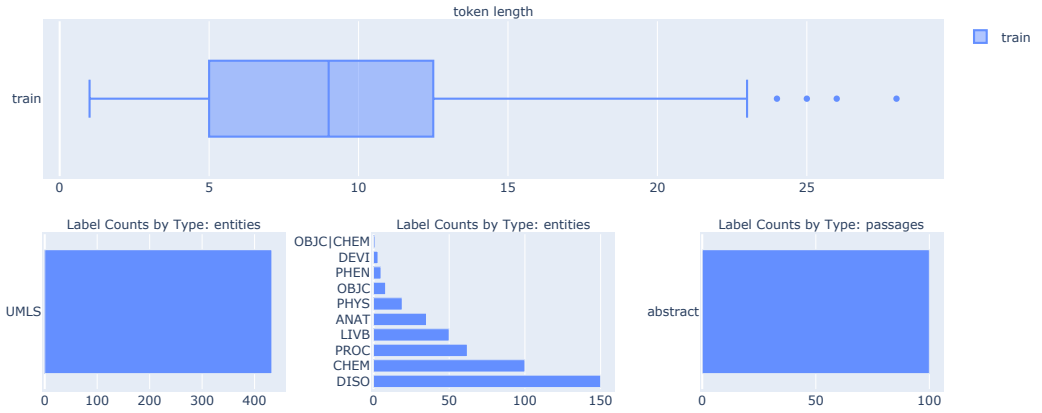

Figure 85: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (Medline English) Data Card

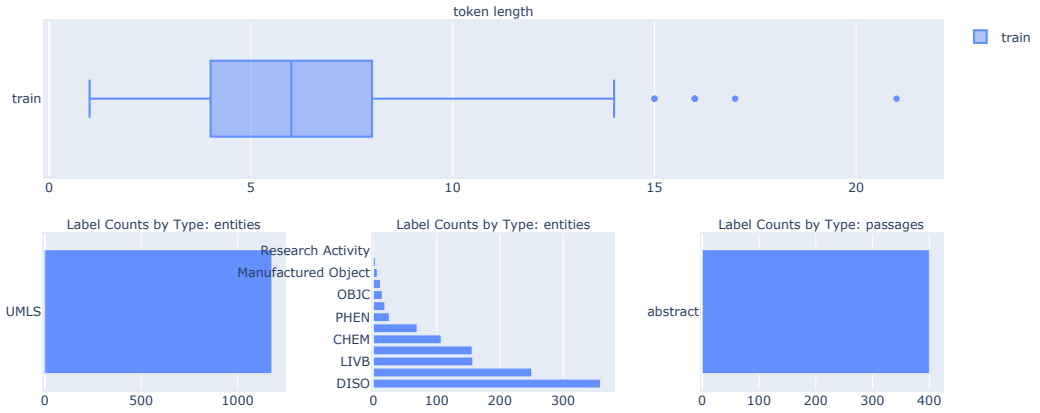

Figure 86: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (Patents English) Data Card

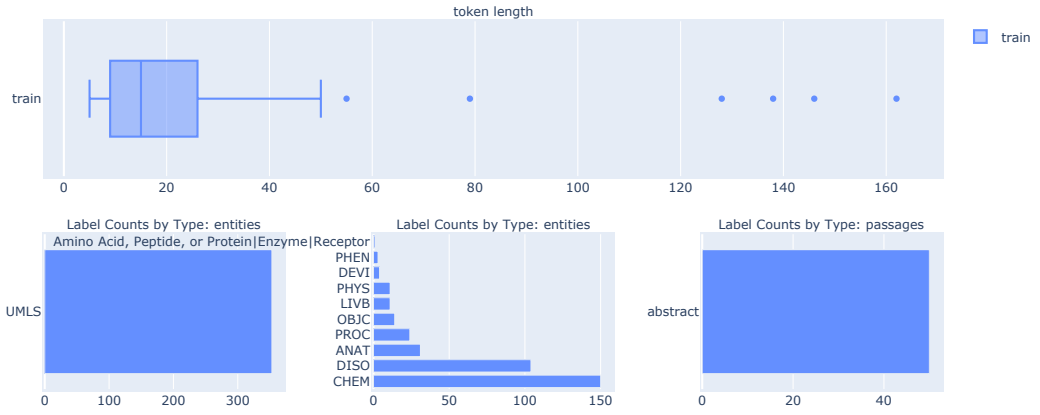

Figure 87: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (EMEA Spanish) Data Card

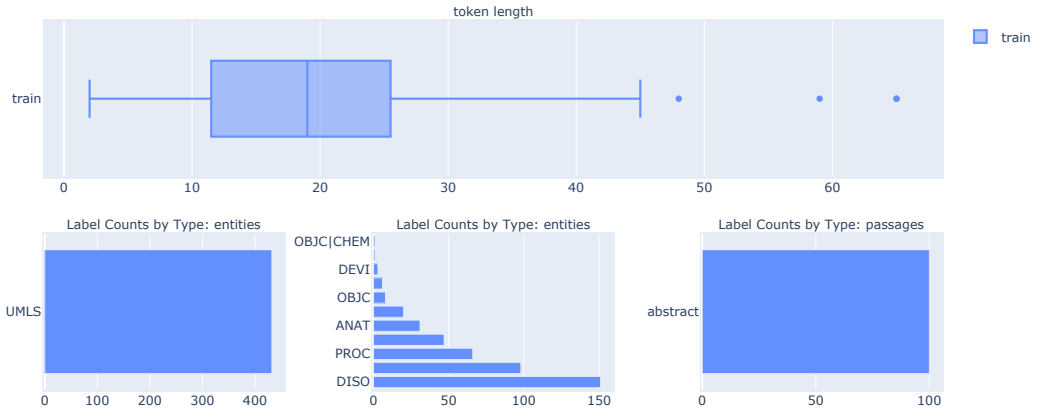

Figure 88: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: Spanish

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (Medline Spanish) Data Card

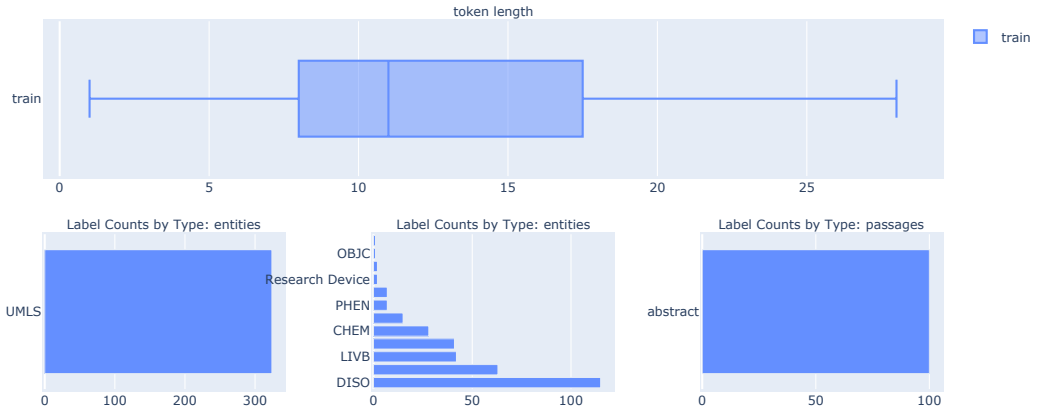

Figure 89: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: Spanish

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (EMEA French) Data Card

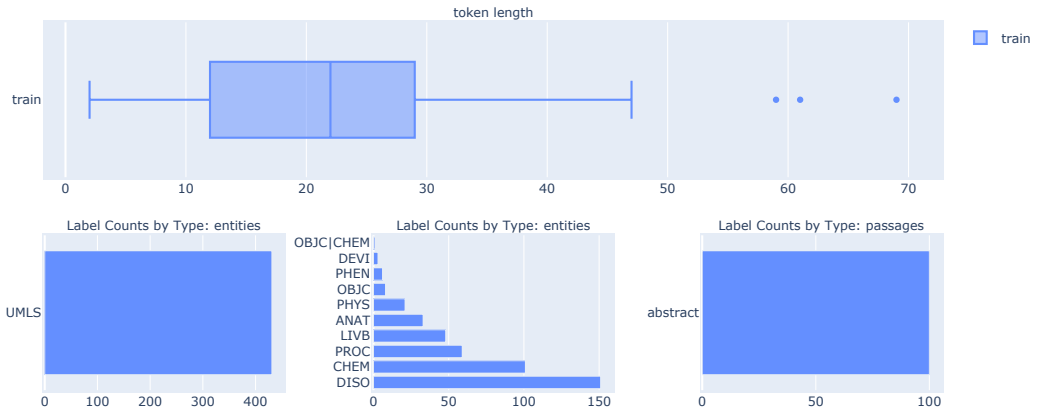

Figure 90: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: French

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (Medline French) Data Card

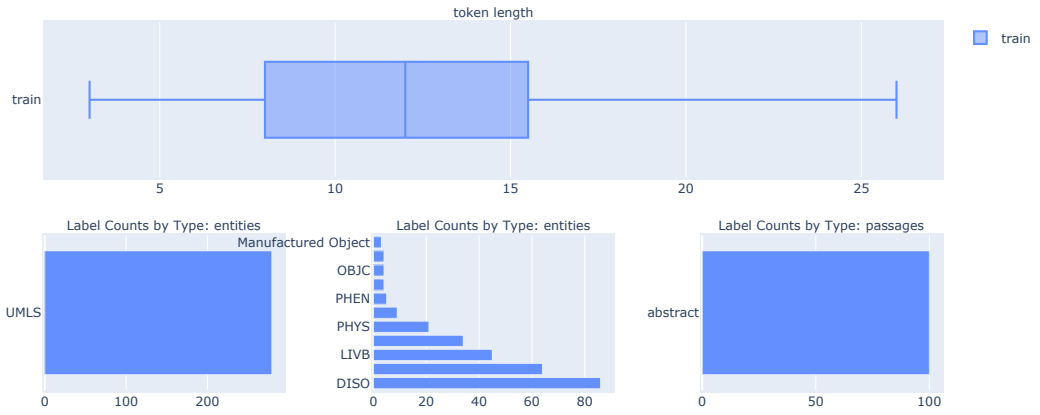

Figure 91: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: French

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (Patents French) Data Card

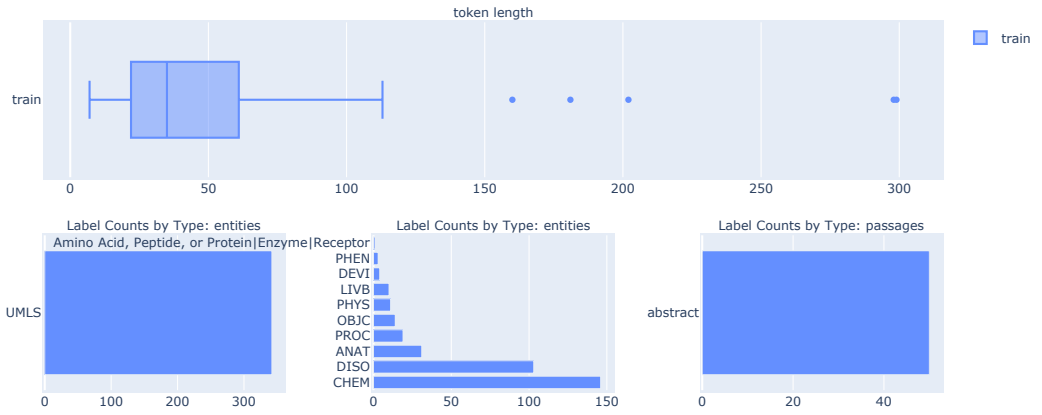

Figure 92: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: French

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Mantra GSC (EMEA Dutch) Data Card

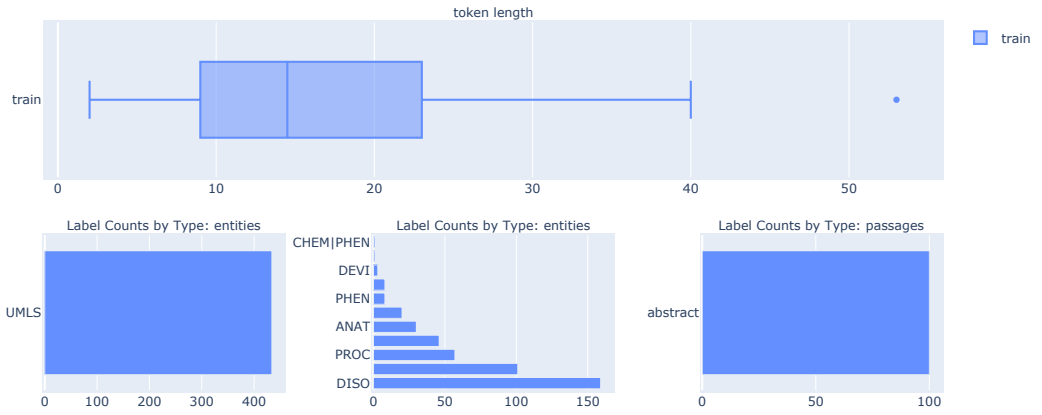

Figure 93: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Tasks: named entity recognition, named entity disambiguation

Languages: Dutch

Schemas: `KB`

Splits: train

Mantra GSC (Medline Dutch) Data Card

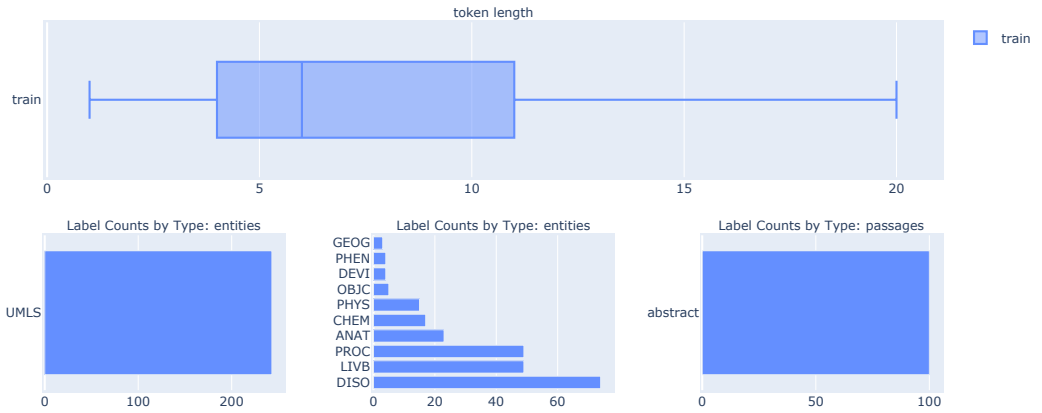

Figure 94: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. **Homepage:** `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

URL: `https://biosemantics.erasmusmc.nl/index.php/resources/mantra-gsc`

Licensing: CC_BY_4p0

Languages: Dutch

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

MayoSRS Data Card

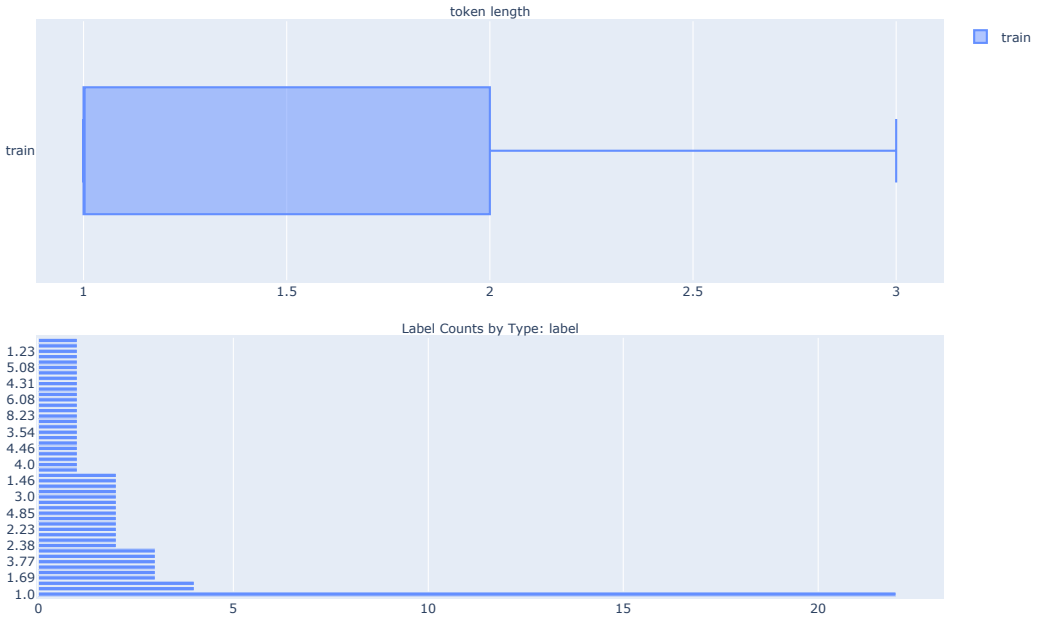

Figure 95: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: MayoSRS consists of 101 clinical term pairs whose relatedness was determined by nine medical coders and three physicians from the Mayo Clinic.

Homepage: `https://conservancy.umn.edu/handle/11299/196265`

URL: `https://conservancy.umn.edu/handle/11299/196265`

Licensing: CC0_1p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

MedQA (English) Data Card

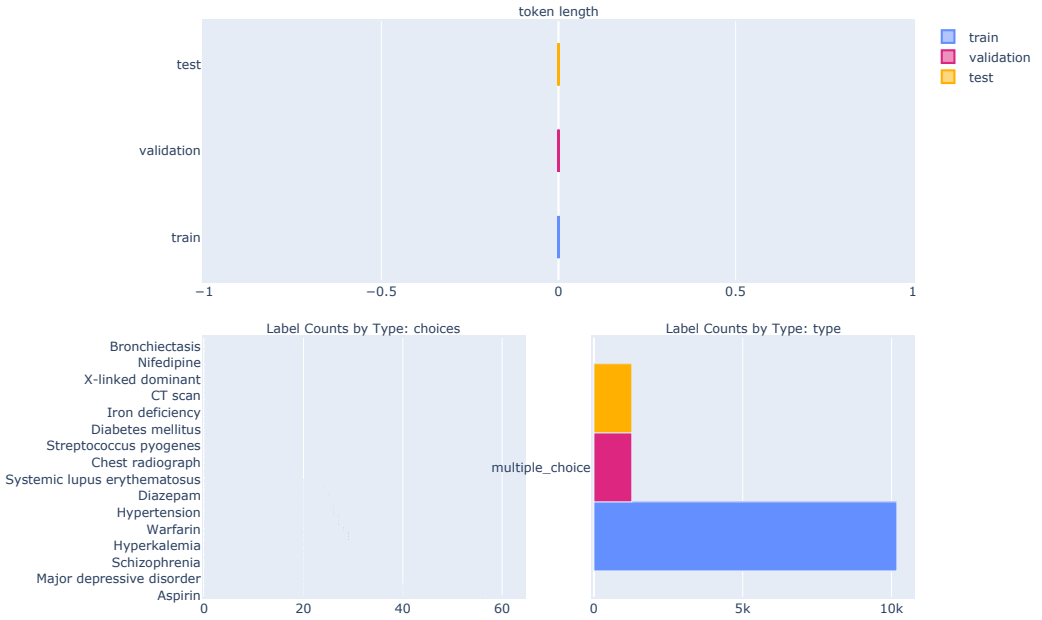

Figure 96: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: In this work, we present the first free-form multiple-choice OpenQA dataset for solving medical problems, MedQA, collected from the professional medical board exams. It covers three languages: English, simplified Chinese, and traditional Chinese, and contains 12,723, 34,251, and 14,123 questions for the three languages, respectively. Together with the question data, we also collect and release a large-scale corpus from medical textbooks from which the reading comprehension models can obtain necessary knowledge for answering the questions.

Homepage: `https://github.com/jind11/MedQA`

URL: `https://github.com/jind11/MedQA`

Licensing: UNKNOWN

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, test, validation

MedQA (Traditional Chinese) Data Card

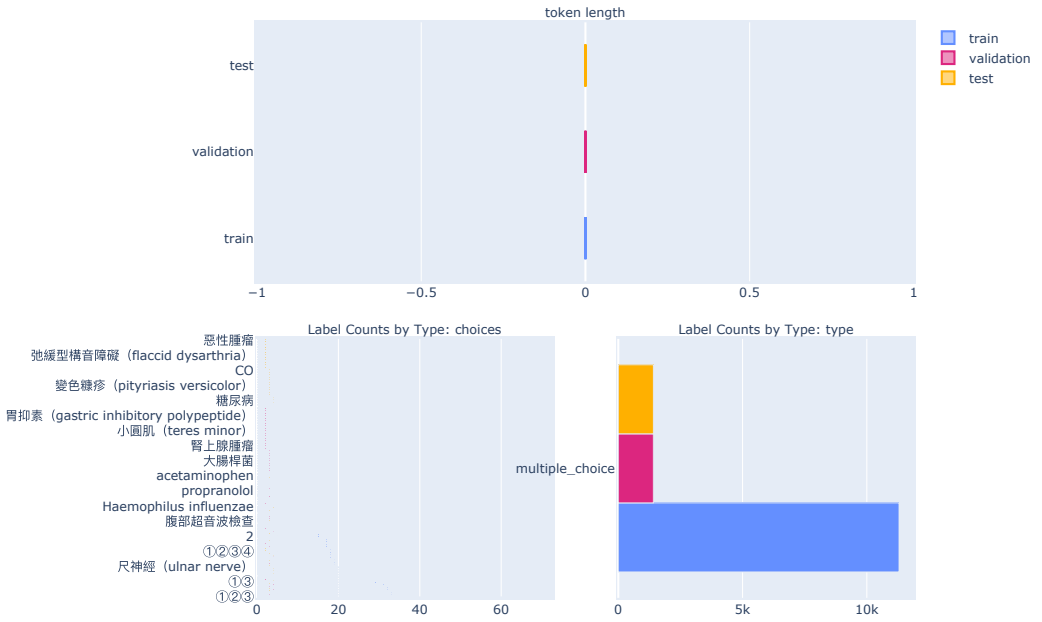

Figure 97: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: In this work, we present the first free-form multiple-choice OpenQA dataset for solving medical problems, MedQA, collected from the professional medical board exams. It covers three languages: English, simplified Chinese, and traditional Chinese, and contains 12,723, 34,251, and 14,123 questions for the three languages, respectively. Together with the question data, we also collect and release a large-scale corpus from medical textbooks from which the reading comprehension models can obtain necessary knowledge for answering the questions.

Homepage: `https://github.com/jind11/MedQA`

URL: `https://github.com/jind11/MedQA`

Licensing: UNKNOWN

Languages: Traditional Chinese

Tasks: question answering

Schemas: `QA`

Splits: train, test, validation

MedQA (Simplified Chinese) Data Card

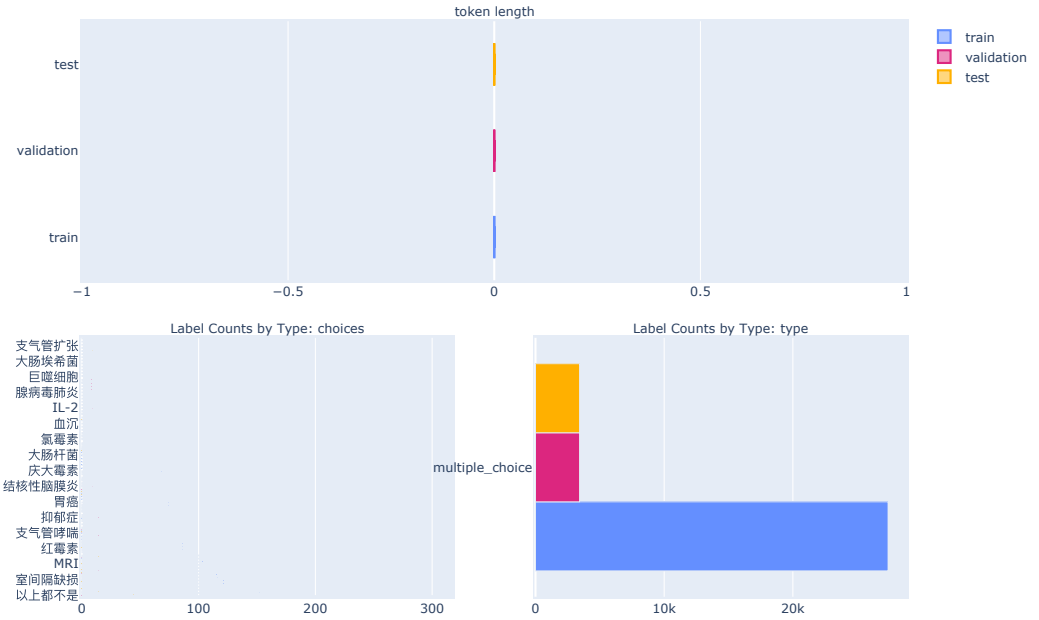

Figure 98: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: In this work, we present the first free-form multiple-choice OpenQA dataset for solving medical problems, MedQA, collected from the professional medical board exams. It covers three languages: English, simplified Chinese, and traditional Chinese, and contains 12,723, 34,251, and 14,123 questions for the three languages, respectively. Together with the question data, we also collect and release a large-scale corpus from medical textbooks from which the reading comprehension models can obtain necessary knowledge for answering the questions.

Homepage: `https://github.com/jind11/MedQA`

URL: `https://github.com/jind11/MedQA`

Licensing: UNKNOWN

Languages: Simplified Chinese

Tasks: question answering

Schemas: `QA`

Splits: train, test, validation

MedDialog (English) Data Card

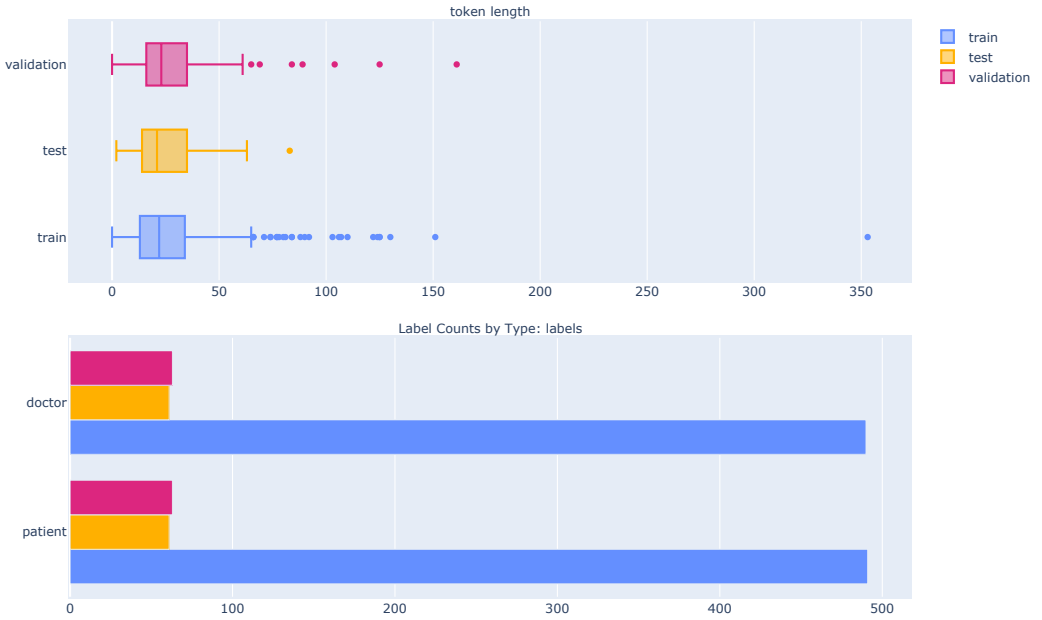

Figure 99: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The MedDialog dataset (English) contains conversations (in English) between doctors and patients.It has 0.26 million dialogues. The data is continuously growing and more dialogues will be added. The raw dialogues are from healthcaremagic.com and icliniq.com.All copyrights of the data belong to healthcaremagic.com and icliniq.com.

Homepage: `https://github.com/UCSD-AI4H/Medical-Dialogue-System`

URL: `https://github.com/UCSD-AI4H/Medical-Dialogue-System`

Licensing: UNKNOWN

Languages: Chinese

Tasks: text classification

Schemas: `TEXT`

Splits: train, validation, test

MedDialog (Chinese) Data Card

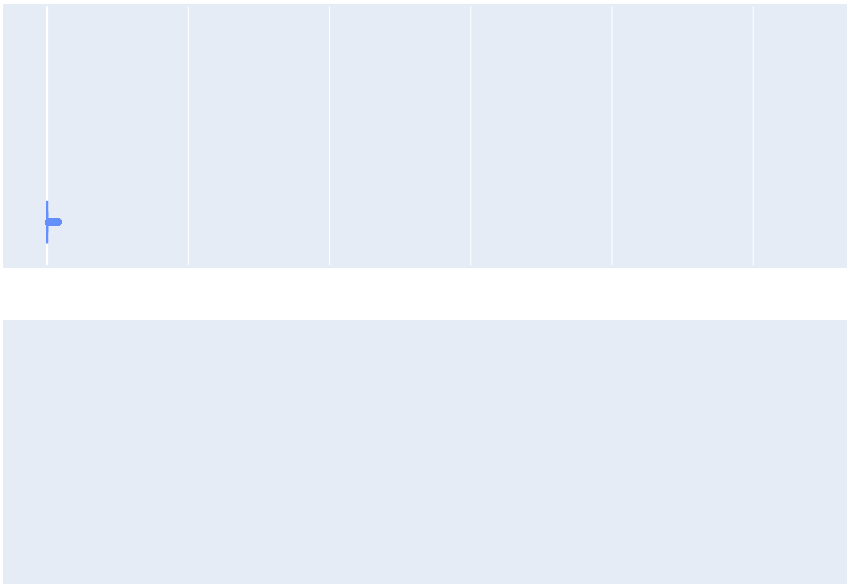

Figure 100: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The MedDialog dataset (English) contains conversations (in English) between doctors and patients.It has 0.26 million dialogues. The data is continuously growing and more dialogues will be added. The raw dialogues are from healthcaremagic.com and icliniq.com.All copyrights of the data belong to healthcaremagic.com and icliniq.com.

Homepage: `https://github.com/UCSD-AI4H/Medical-Dialogue-System`

URL: `https://github.com/UCSD-AI4H/Medical-Dialogue-System`

Licensing: UNKNOWN

Languages: Chinese

Tasks: text classification

Schemas: `TEXT`

Splits: train, validation, test

MEDDOCAN Data Card

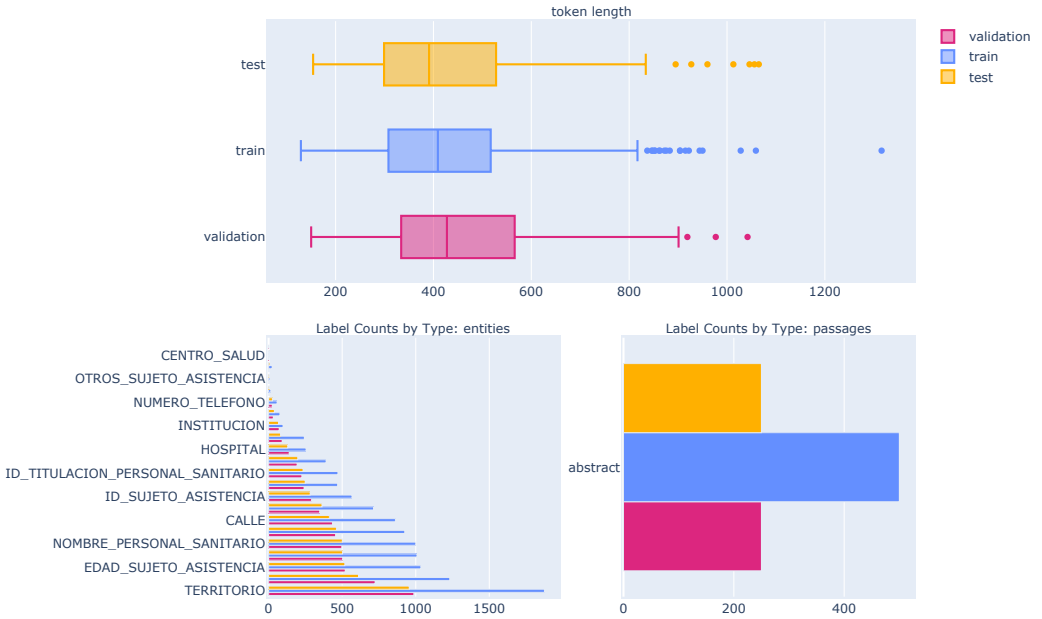

Figure 101: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: MEDDOCAN: Medical Document Anonymization Track This dataset is designed for the MEDDOCAN task, sponsored by Plan de Impulso de las Tecnologías del Lenguaje. It is a manually classified collection of 1,000 clinical case reports derived from the Spanish Clinical Case Corpus (SPACCC), enriched with PHI expressions.The annotation of the entire set of entity mentions was carried out by experts annotators and it includes 29 entity types relevant for the annonymiation of medical documents.22 of these annotation types are actually present in the corpus: TERRITORIO, FECHAS, EDAD_SUJETO_ASISTENCIA, NOMBRE_SUJETO_ASISTENCIA, NOMBRE_PERSONAL_SANITARIO, SEXO_SUJETO_ASISTENCIA, CALLE, PAIS, ID_SUJETO_ASISTENCIA, CORREO, ID_TITULACION_PERSONAL_SANITARIO,ID_ASEGURAMIENTO, HOSPITAL, FAMILIARES_SUJETO_ASISTENCIA, INSTITUCION, ID_CONTACTO ASISTENCIAL,NUMERO_TELEFONO, PROFESION, NUMERO_FAX, OTROS_SUJETO_ASISTENCIA, CENTRO_SALUD, ID_EMPLEO_PERSONAL_SANITARIO For further information, please visit `https://temu.bsc.es/meddocan/` or send an email to encargo-pln-life@bsc.es

Homepage: `https://temu.bsc.es/meddocan/`

URL: `https://temu.bsc.es/meddocan/`

Licensing: CC_BY_4p0

Languages: Spanish

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

MedHop Data Card

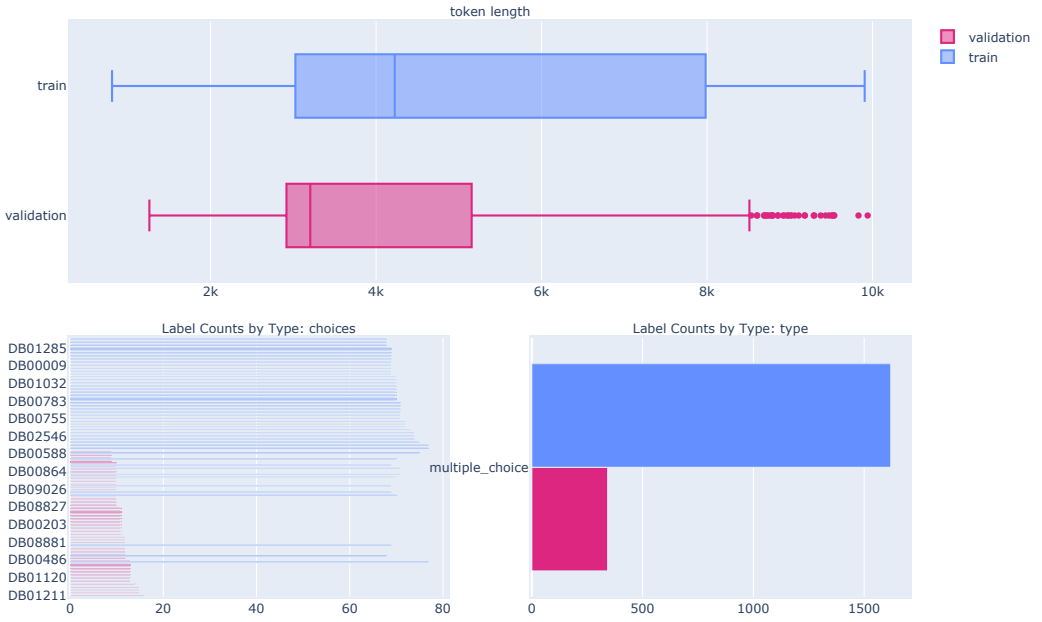

Figure 102: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: With the same format as WikiHop, this dataset is based on research paper abstracts from PubMed, and the queries are about interactions between pairs of drugs. The correct answer has to be inferred by combining information from a chain of reactions of drugs and proteins.

Homepage: `http://qangaroo.cs.ucl.ac.uk/`

URL: `http://qangaroo.cs.ucl.ac.uk/`

Licensing: CC_BY_SA_3p0

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, validation

MEDIQA QA Data Card

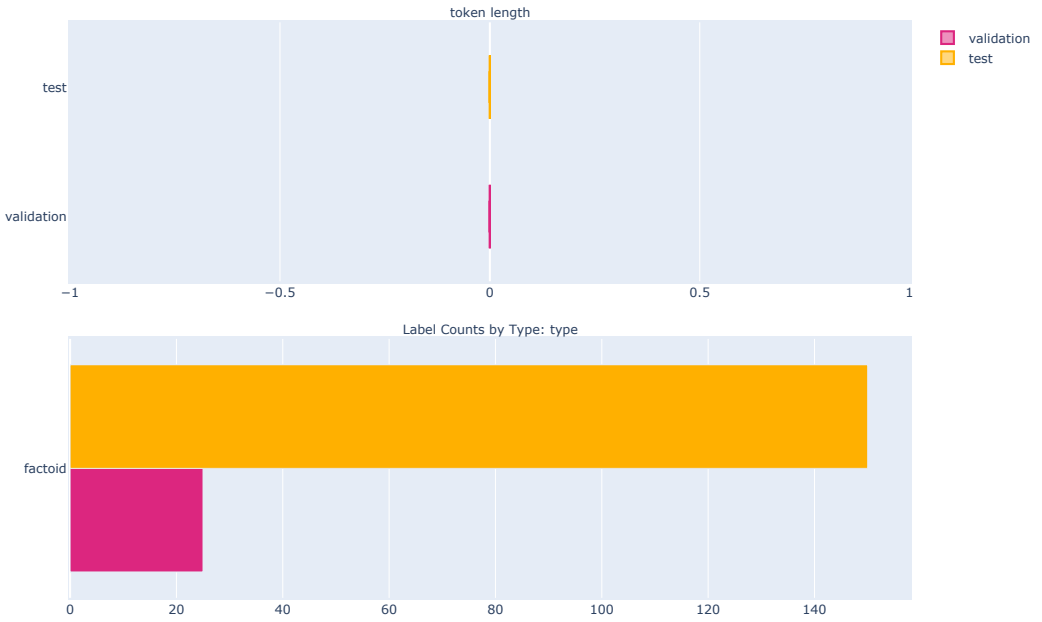

Figure 103: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The MEDIQA challenge is an ACL-BioNLP 2019 shared task aiming to attract further research efforts in Natural Language Inference (NLI), Recognizing Question Entailment (RQE), and their applications in medical Question Answering (QA).Mailing List: `https://groups.google.com/forum/#!forum/bionlp-mediqa`.

In the QA task, participants are tasked to:- filter/classify the provided answers (1: correct, 0: incorrect).- re-rank the answers.

Homepage: `https://sites.google.com/view/mediqa2019`

URL: `https://sites.google.com/view/mediqa2019`

Licensing: UNKNOWN

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train_live_qa_med, train_alexa, validation, test

MEDIQA RQE Data Card

Figure 104: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The MEDIQA challenge is an ACL-BioNLP 2019 shared task aiming to attract further research efforts in Natural Language Inference (NLI), Recognizing Question Entailment (RQE), and their applications in medical Question Answering (QA).Mailing List: `https://groups.google.com/forum/#!forum/bionlp-mediqa`.

In the QA task, participants are tasked to:- filter/classify the provided answers (1: correct, 0: incorrect).- re-rank the answers.

Homepage: `https://sites.google.com/view/mediqa2019`

URL: `https://sites.google.com/view/mediqa2019`

Licensing: UNKNOWN

Languages: English

Tasks: textual entailment

Schemas: `TE`

Splits: train, validation, test

MedMentions (Full) Data Card

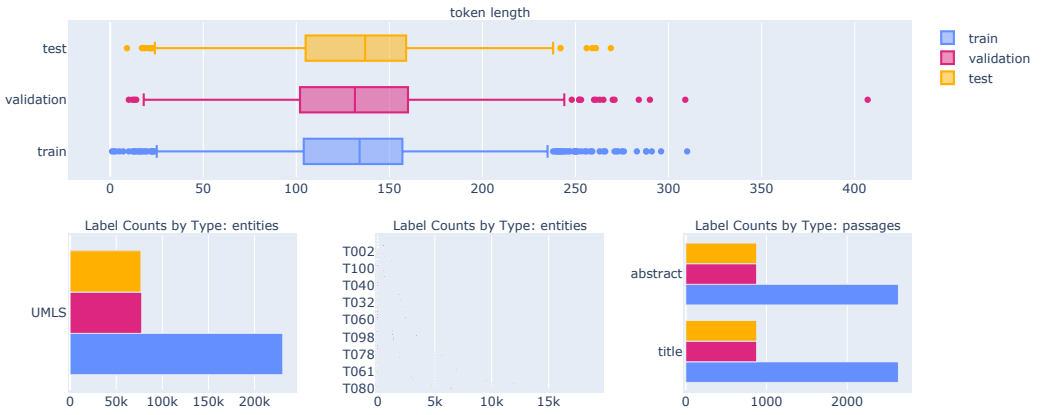

Figure 105: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: MedMentions is a new manually annotated resource for the recognition of biomedical concepts.What distinguishes MedMentions from other annotated biomedical corpora is its size (over 4,000abstracts and over 350,000 linked mentions), as well as the size of the concept ontology (over3 million concepts from UMLS 2017) and its broad coverage of biomedical disciplines.Corpus: The MedMentions corpus consists of 4,392 papers (Titles and Abstracts) randomly selected from among papers released on PubMed in 2016, that were in the biomedical field, published inthe English language, and had both a Title and an Abstract.

Annotators: We recruited a team of professional annotators with rich experience in biomedical content curation to exhaustively annotate all UMLS® (2017AA full version) entity mentions in these papers.Annotation quality: We did not collect stringent IAA (Inter-annotator agreement) data. To gain insight on the annotation quality of MedMentions, we randomly selected eight papers from the annotated corpus, containing a total of 469 concepts. Two biologists ('Reviewer') who did not participate in the annotation task then each reviewed four papers. The agreement between Reviewers and Annotators, an estimate of the Precision of the annotations, was 97.3

Homepage: `https://github.com/chanzuckerberg/MedMentions`

URL: `https://github.com/chanzuckerberg/MedMentions`

Licensing: CC0_1p0

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train, test, validation

MedMentions (ST21PV) Data Card

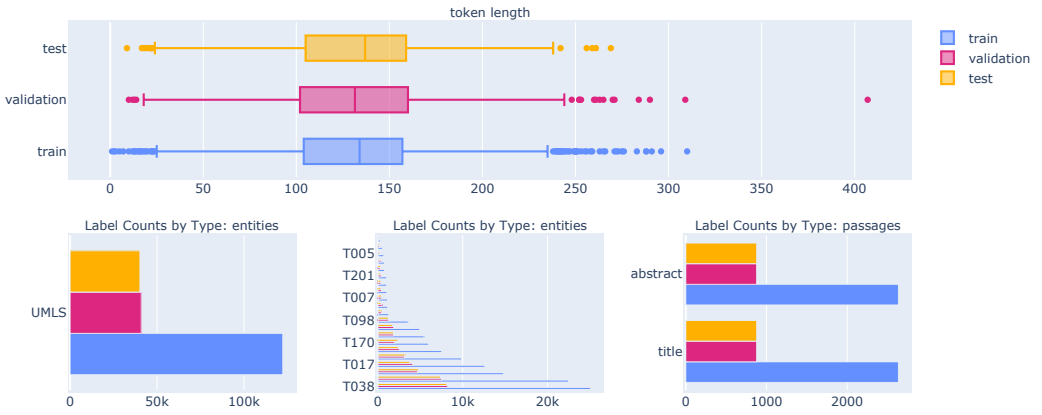

Figure 106: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: MedMentions is a new manually annotated resource for the recognition of biomedical concepts.What distinguishes MedMentions from other annotated biomedical corpora is its size (over 4,000abstracts and over 350,000 linked mentions), as well as the size of the concept ontology (over3 million concepts from UMLS 2017) and its broad coverage of biomedical disciplines.Corpus: The MedMentions corpus consists of 4,392 papers (Titles and Abstracts) randomly selected from among papers released on PubMed in 2016, that were in the biomedical field, published in the English language, and had both a Title and an Abstract.

Annotators: We recruited a team of professional annotators with rich experience in biomedical content curation to exhaustively annotate all UMLS® (2017AA full version) entity mentions in these papers.Annotation quality: We did not collect stringent IAA (Inter-annotator agreement) data. To gain insight on the annotation quality of MedMentions, we randomly selected eight papers from the annotated corpus, containing a total of 469 concepts. Two biologists ('Reviewer') who did not participate in the annotation task then each reviewed four papers. The agreement between Reviewers and Annotators, an estimate of the Precision of the annotations, was 97.3**Homepage:** `https://github.com/chanzuckerberg/MedMentions`

URL: `https://github.com/chanzuckerberg/MedMentions`

Licensing: CC0_1p0

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train, test, validation

MeQSUM Data Card

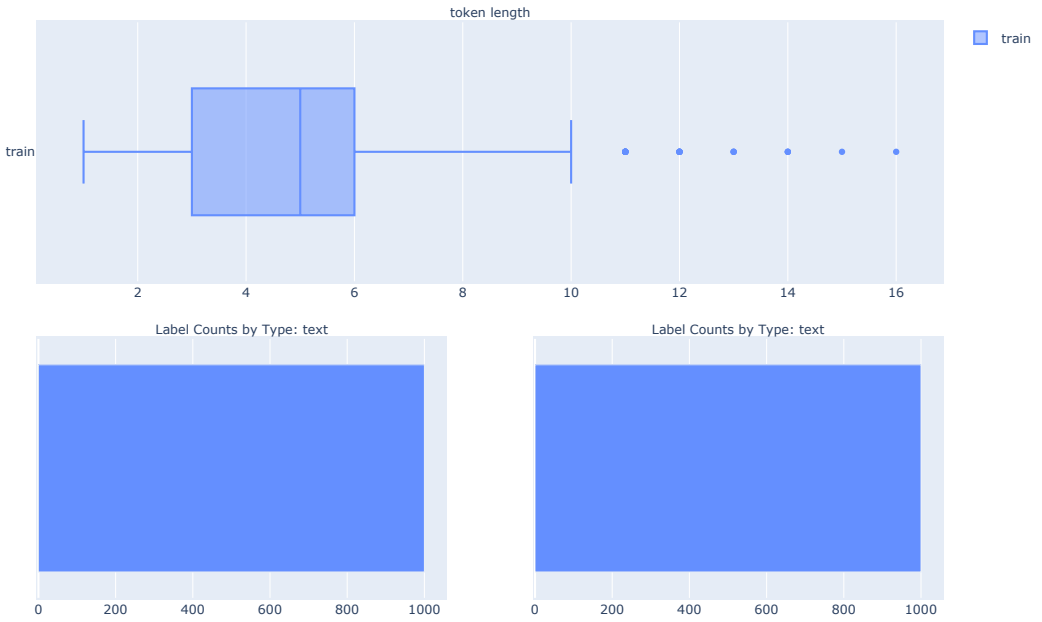

Figure 107: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Dataset for medical question summarization introduced in the ACL 2019 paper "On the Summarization of Consumer Health Questions". Question understanding is one of the main challenges in question answering. In real world applications,users often submit natural language questions that are longer than needed and include peripheral information that increases the complexity of the question, leading to substantially more false positives in answer retrieval. In this paper, we study neural abstractive models for medical question summarization. We introduce the MeQSum corpus of 1,000summarized consumer health questions. **Homepage:** `https://github.com/abachaa/MeQSum`

URL: `https://github.com/abachaa/MeQSum`

Licensing: UNKNOWN

Languages: English

Tasks: summarization

Schemas: `T2T`

Splits: train

MiniMayoSRS Data Card

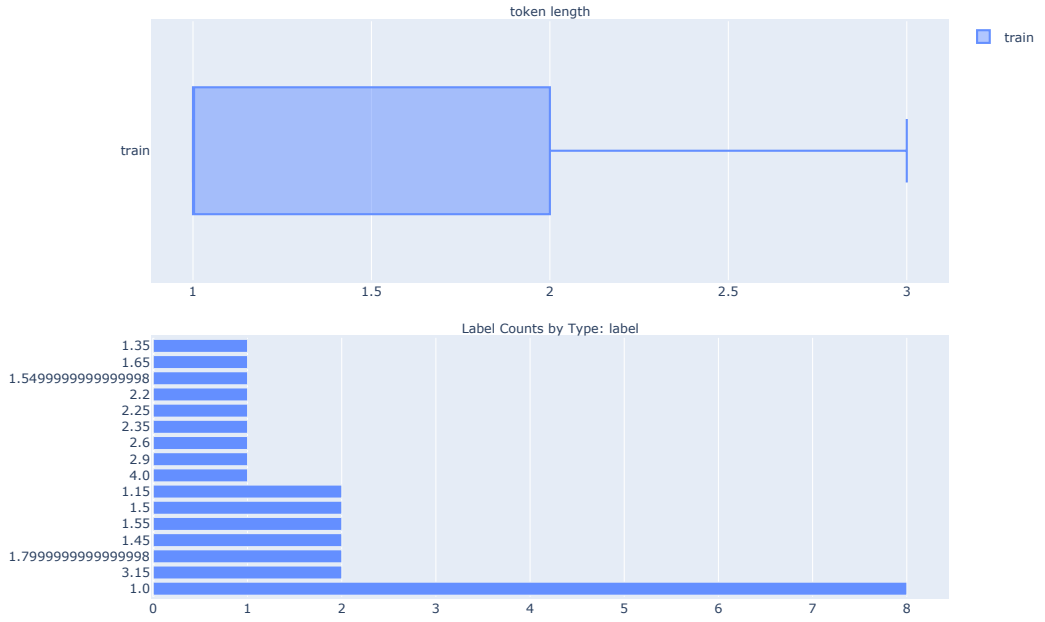

Figure 108: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: MiniMayoSRS is a subset of the MayoSRS and consists of 30 term pairs on which a higher inter-annotator agreement was achieved. The average correlation between physicians is 0.68. The average correlation between medical coders is 0.78.

Homepage: `https://conservancy.umn.edu/handle/11299/196265`

URL: `https://conservancy.umn.edu/handle/11299/196265`

Licensing: CC0_1p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

miRNA Data Card

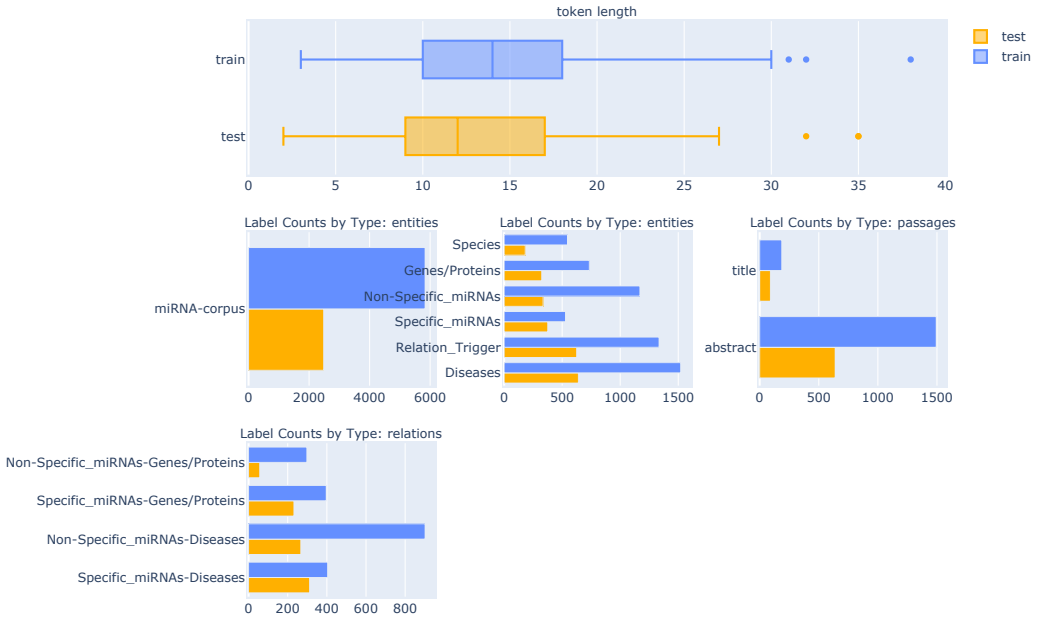

Figure 109: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The corpus consists of 301 Medline citations. The documents were screened for mentions of miRNA in the abstract text. Gene, disease and miRNA entities were manually annotated. The corpus comprises of two separate files, a train and a test set, coming from 201 and 100 documents respectively.

Homepage: `https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/download-mirna-test-corpus.html`

URL: `https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/download-mirna-test-corpus.html`

Licensing: CC_BY_NC_3p0

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train, test

MLEE Data Card

Figure 110: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: MLEE is an event extraction corpus consisting of manually annotated abstracts of paperson angiogenesis. It contains annotations for entities, relations, events and coreferences. The annotations span molecular, cellular, tissue, and organ-level processes.

Homepage: `http://www.nactem.ac.uk/MLEE/`

URL: `http://www.nactem.ac.uk/MLEE/`

Licensing: CC_BY_NC_SA_3p0

Languages: English

Tasks: named entity recognition, relation extraction, event extraction, coreference resolution

Schemas: `KB`

Splits: train, validation, test

MQP Data Card

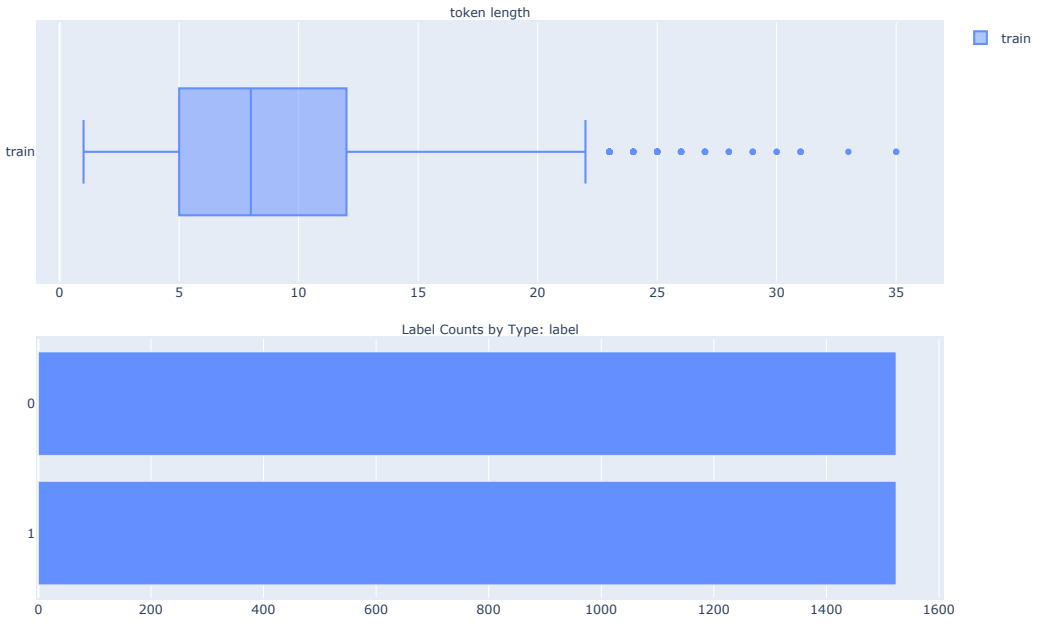

Figure 111: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Medical Question Pairs dataset by McCreery et al (2020) contains pairs of medical questions and paraphrased versions of the question prepared by medical professional. Paraphrased versions were labelled as similar (syntactically dissimilar but contextually similar) or dissimilar (syntactically may look similar but contextually dissimilar). Labels 1: similar, 0: dissimilar

Homepage: `https://github.com/curai/medical-question-pair-dataset`

URL: `https://github.com/curai/medical-question-pair-dataset`

Licensing: UNKNOWN

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

MuchMore Data Card

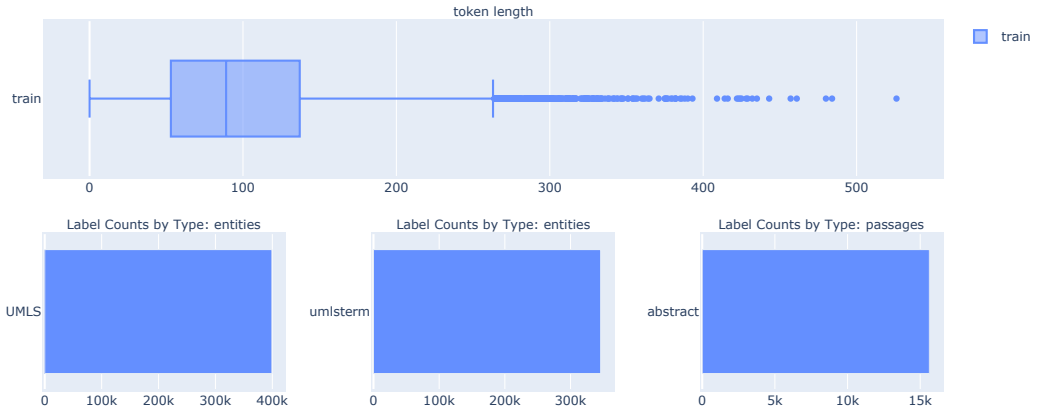

Figure 112: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The corpus used in the MuchMore project is a parallel corpus of English-German scientific medical abstracts obtained from the Springer Link web site. The corpus consists approximately of 1 million tokens for each language. Abstracts are from 41 medical journals, each of which constitutes a relatively homogeneous medical sub-domain (e.g.Neurology, Radiology, etc.). The corpus of downloaded HTML documents is normalized invarious ways, in order to produce a clean, plain text version, consisting of a title, abstract and keywords. Additionally, the corpus was aligned on the sentence level.

Automatic (!) annotation includes: Part-of-Speech; Morphology (inflection anddecomposition); Chunks; Semantic Classes (UMLS: Unified Medical Language System, MeSH: Medical Subject Headings, EuroWordNet); Semantic Relations from UMLS.

Homepage: `https://muchmore.dfki.de/resources1.htm`

URL: `https://muchmore.dfki.de/resources1.htm`

Licensing: UNKNOWN

Languages: English German

Tasks: named entity recognition

Schemas: `KB`

Splits: train

MuchMore (Translation) Data Card

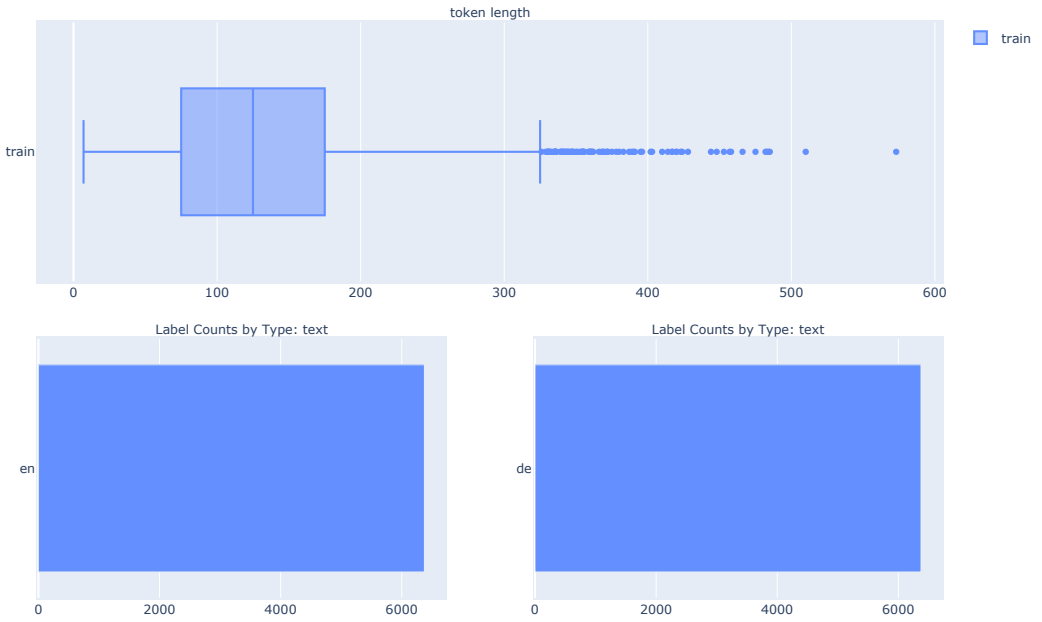

Figure 113: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The corpus used in the MuchMore project is a parallel corpus of English-German scientific medical abstracts obtained from the Springer Link web site. The corpus consists approximately of 1 million tokens for each language. Abstracts are from 41 medical journals, each of which constitutes a relatively homogeneous medical sub-domain (e.g.Neurology, Radiology, etc.). The corpus of downloaded HTML documents is normalized invarious ways, in order to produce a clean, plain text version, consisting of a title, abstract and keywords. Additionally, the corpus was aligned on the sentence level.

Automatic (!) annotation includes: Part-of-Speech; Morphology (inflection anddecomposition); Chunks; Semantic Classes (UMLS: Unified Medical Language System, MeSH: Medical Subject Headings, EuroWordNet); Semantic Relations from UMLS.

Homepage: `https://muchmore.dfki.de/resources1.htm`

URL: `https://muchmore.dfki.de/resources1.htm`

Licensing: UNKNOWN

Languages: English, German

Tasks: translation

Schemas: `T2T KB`

Splits: train

Multi-XScience Data Card

Figure 114: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Multi-document summarization is a challenging task for which there exists little large-scale datasets. We propose Multi-XScience, a large-scale multi-document summarization dataset created from scientific articles. Multi-XScience introduces a challenging multi-document summarization task: writing the related-work section of a paper based on its abstract and the articles it references. Our work is inspired by extreme summarization, a dataset construction protocol that favours abstractive modeling approaches. Descriptive statistics and empirical results—using several state-of-the-art models trained on the Multi-XScience dataset—reveal that Multi-XScience is well suited for abstractive models.

Homepage: `https://github.com/yaolu/Multi-XScience`

URL: `https://github.com/yaolu/Multi-XScience`

Licensing: MIT

Languages: English

Tasks: summarization, paraphrasing

Schemas: `T2T`

Splits: train, test, validation

MutationFinder Data Card

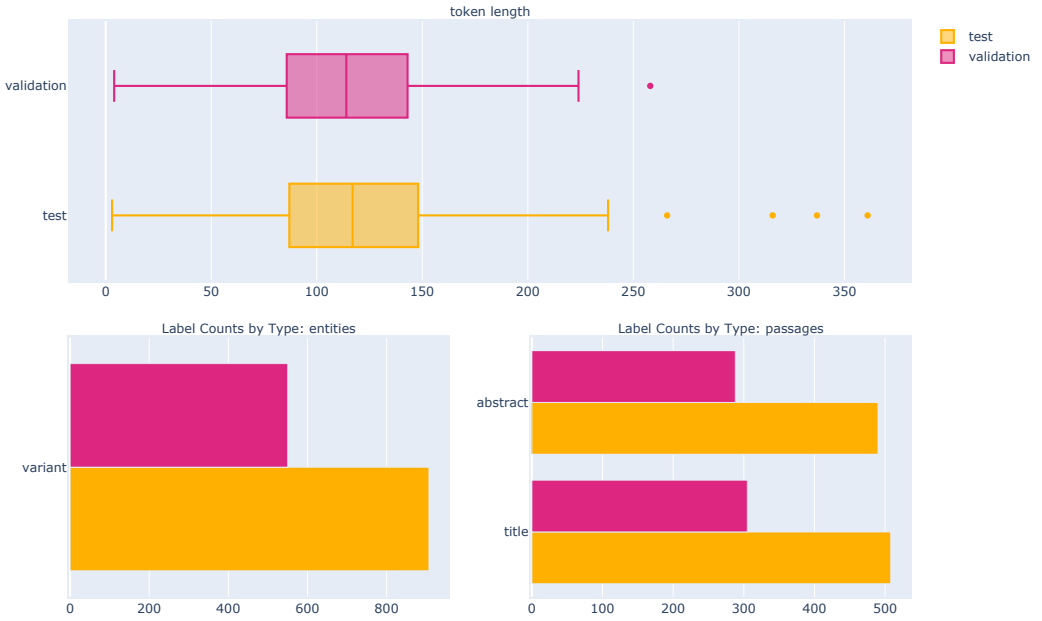

Figure 115: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Gold standard corpus for mutation extraction systems consisting of 1515 human-annotated mutation mentions in 813MEDLINE abstracts. This corpus is divided into development and test subsets. Inter-annotator agreement on this corpus,judged on fifty abstracts, was 94%.

Homepage: `http://mutationfinder.sourceforge.net/`

URL: `http://mutationfinder.sourceforge.net/`

Licensing: Custom license

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: validation, test

NCBI Disease Data Card

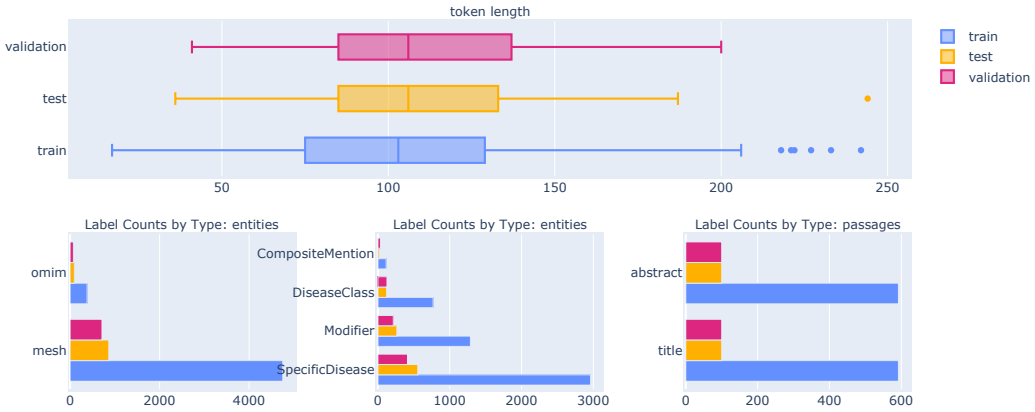

Figure 116: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The NCBI disease corpus is fully annotated at the mention and concept level to serve as a research resource for the biomedical natural language processing community.

Homepage: `https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/`

URL: `https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/`

Licensing: CC0_1p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train, test, validation

NLM-Gene Data Card

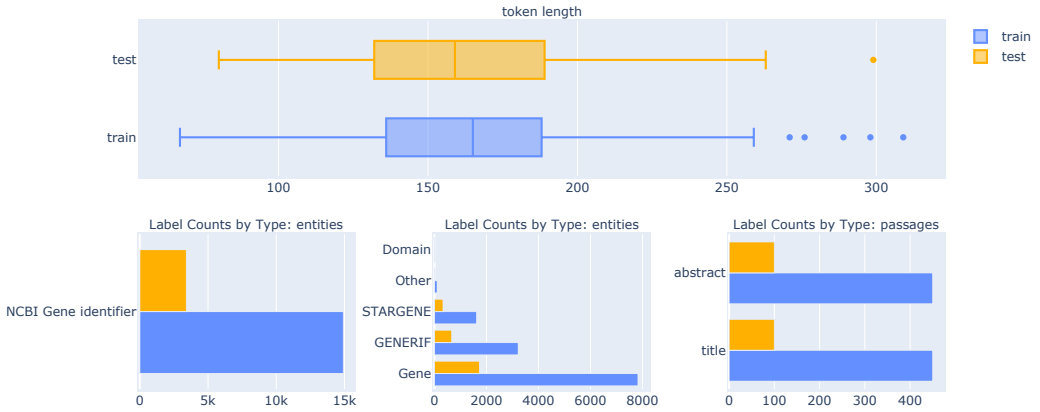

Figure 117: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: NLM-Gene consists of 550 PubMed articles, from 156 journals, and contains more than 15 thousand unique gene names,corresponding to more than five thousand gene identifiers (NCBI Gene taxonomy). This corpus contains gene annotation data from 28 organisms. The annotated articles contain on average 29 gene names, and 10 gene identifiers per article.These characteristics demonstrate that this article set is an important benchmark dataset to test the accuracy of gene recognition algorithms both on multi-species and ambiguous data. The NLM-Gene corpus will be invaluable for advancing text-mining techniques for gene identification tasks in biomedical text.

Homepage: `https://zenodo.org/record/5089049`

URL: `https://zenodo.org/record/5089049`

Licensing: CC0_1p0

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: train, test

NLM-Chem Data Card

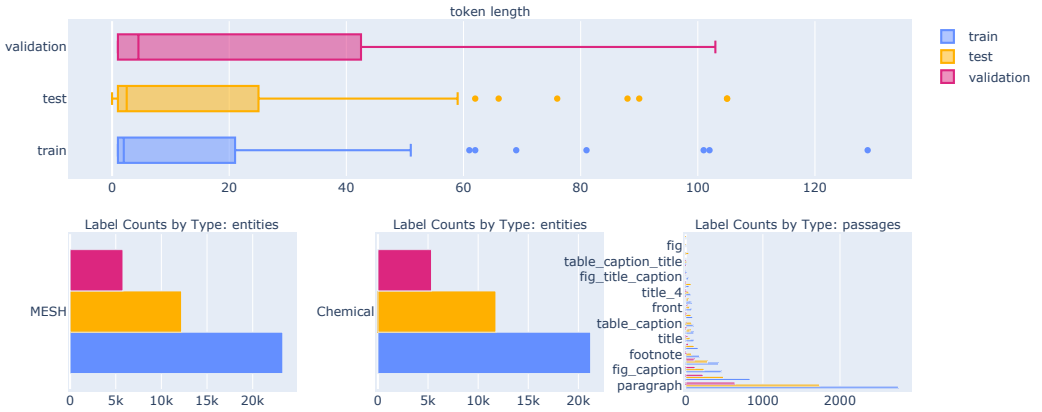

Figure 118: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: NLM-Chem corpus consists of 150 full-text articles from the PubMed Central Open Access dataset,comprising 67 different chemical journals, aiming to cover a general distribution of usage of chemical names in the biomedical literature.Articles were selected so that human annotation was most valuable (meaning that they were rich in bio-entities,and current state-of-the-art named entity recognition systems disagreed on bio-entity recognition.

Homepage: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-2`

URL: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-2`

Licensing: CC0_1p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train, test, validation

NLM-Chem Data Card

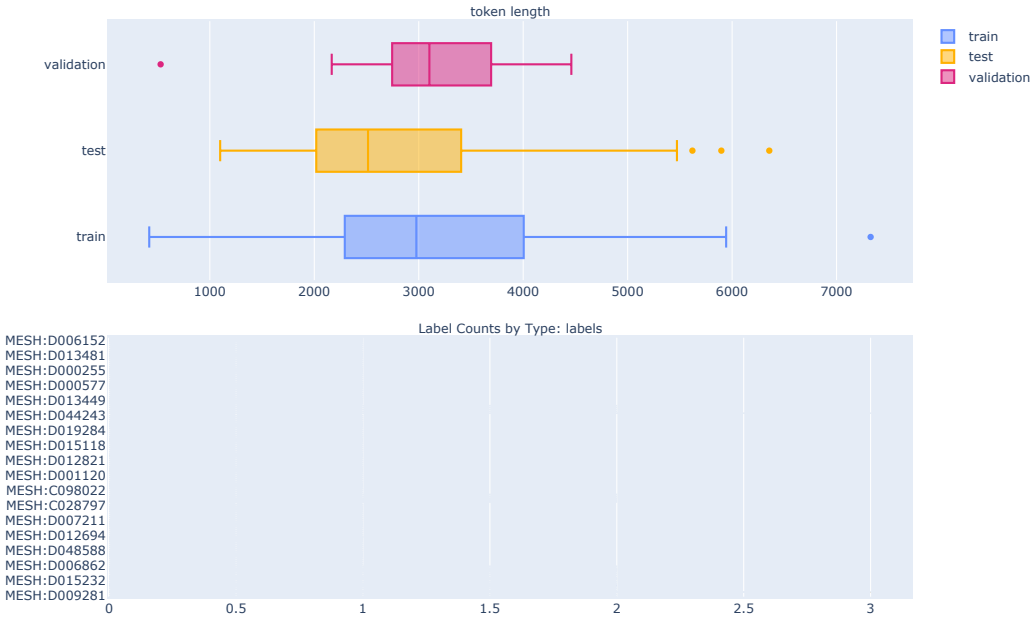

Figure 119: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: NLM-Chem corpus consists of 150 full-text articles from the PubMed Central Open Access dataset,comprising 67 different chemical journals, aiming to cover a general distribution of usage of chemical names in the biomedical literature.Articles were selected so that human annotation was most valuable (meaning that they were rich in bio-entities,and current state-of-the-art named entity recognition systems disagreed on bio-entity recognition.

Homepage: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-2`

URL: `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-2`

Licensing: CC0_1p0

Languages: English

Tasks: text classification

Schemas: `TEXT KB`

Splits: train, test, validation

OSIRIS Data Card

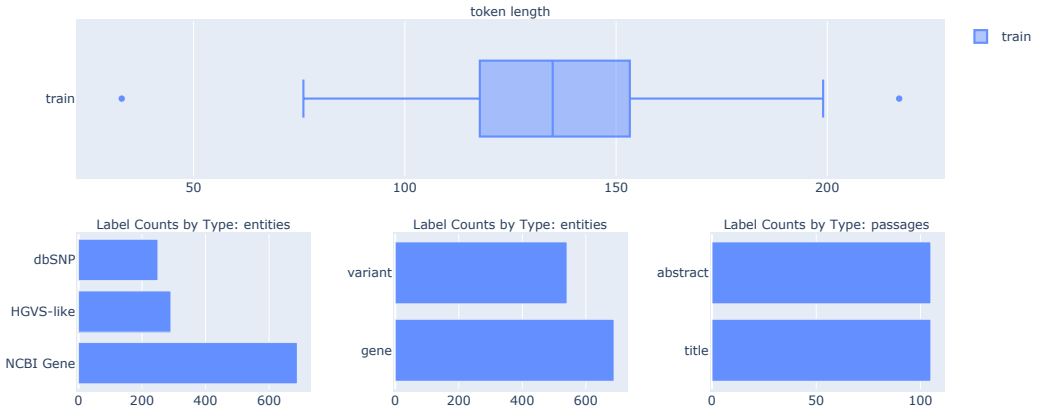

Figure 120: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The OSIRIS corpus is a set of MEDLINE abstracts manually annotatedwith human variation mentions. The corpus is distributed under the terms of the Creative Commons Attribution License Creative Commons Attribution 3.0 Unported License,which permits unrestricted use, distribution, and reproduction in any medium,provided the original work is properly cited (Furlong et al, BMC Bioinformatics 2008, 9:84).

Homepage: `https://sites.google.com/site/laurafurlongweb/databases-and-tools/corpora/`

URL: `https://sites.google.com/site/laurafurlongweb/databases-and-tools/corpora/`

Licensing: CC_BY_3p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

ParaMed Data Card

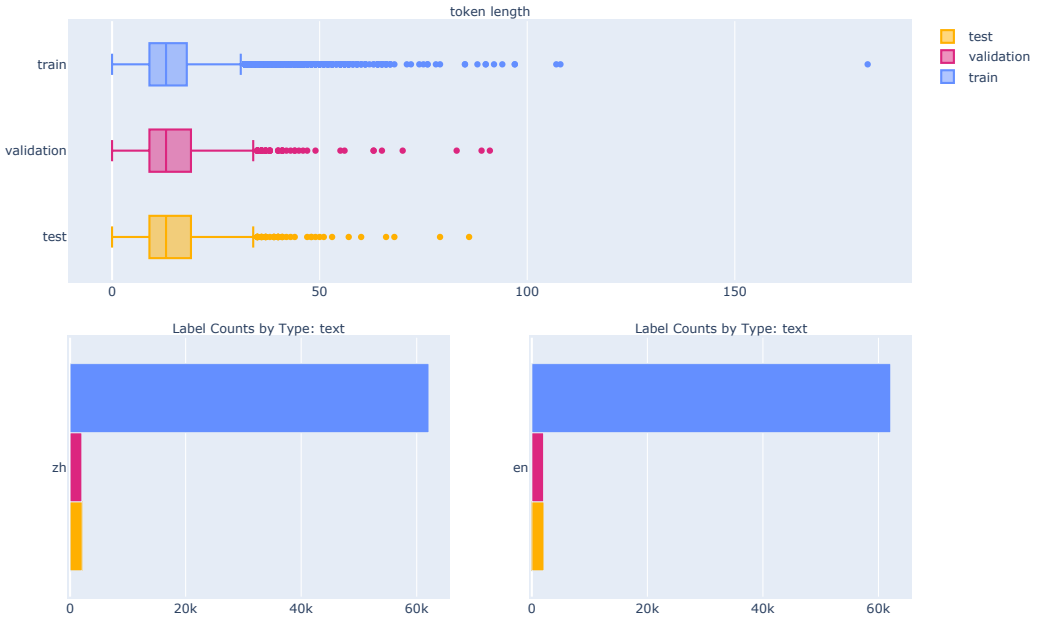

Figure 121: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: NEJM is a Chinese-English parallel corpus crawled from the New England Journal of Medicine website. English articles are distributed through https://www.nejm.org/ and Chinese articles are distributed through http://nejmqianyan.cn/. The corpus contains all article pairs (around 2000 pairs) since 2011.

Homepage: `https://github.com/boxiangliu/ParaMed`

URL: `https://github.com/boxiangliu/ParaMed`

Licensing: CC_BY_4p0

Languages: English Chinese

Tasks: translation

Schemas: `T2T`

Splits: train, validation, test

PDR Data Card

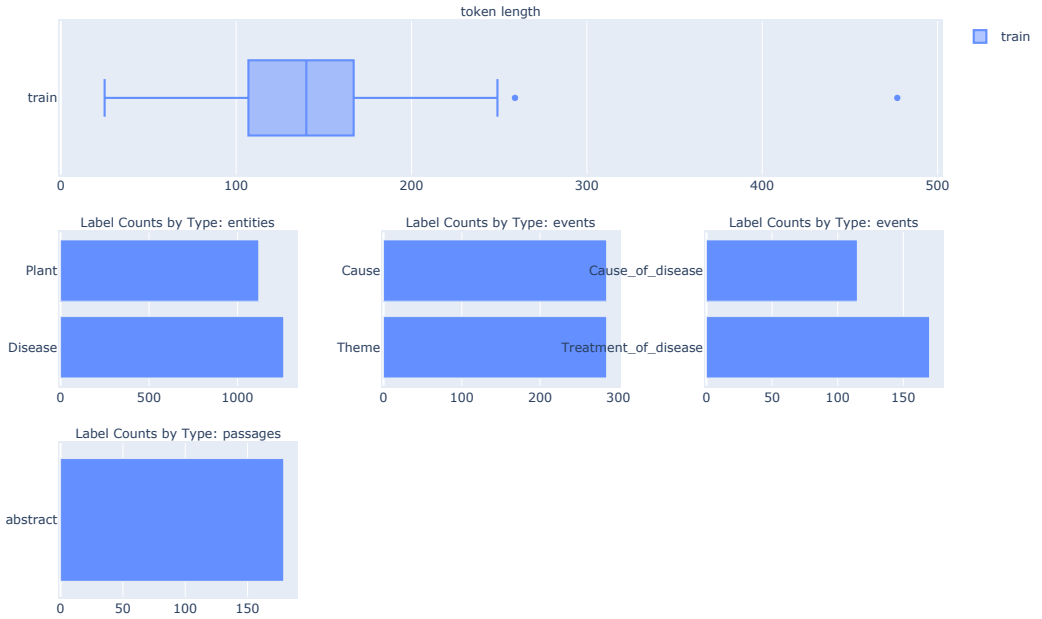

Figure 122: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The corpus of plant-disease relation consists of plants and diseases and their relation to PubMed abstract.The corpus consists of about 2400 plant and disease entities and 300 annotated relations from 179 abstracts.

Homepage: `http://gcancer.org/pdr/`

URL: `http://gcancer.org/pdr/`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition, coreference resolution, event extraction

Schemas: `KB`

Splits: train

PharmaCoNER Data Card

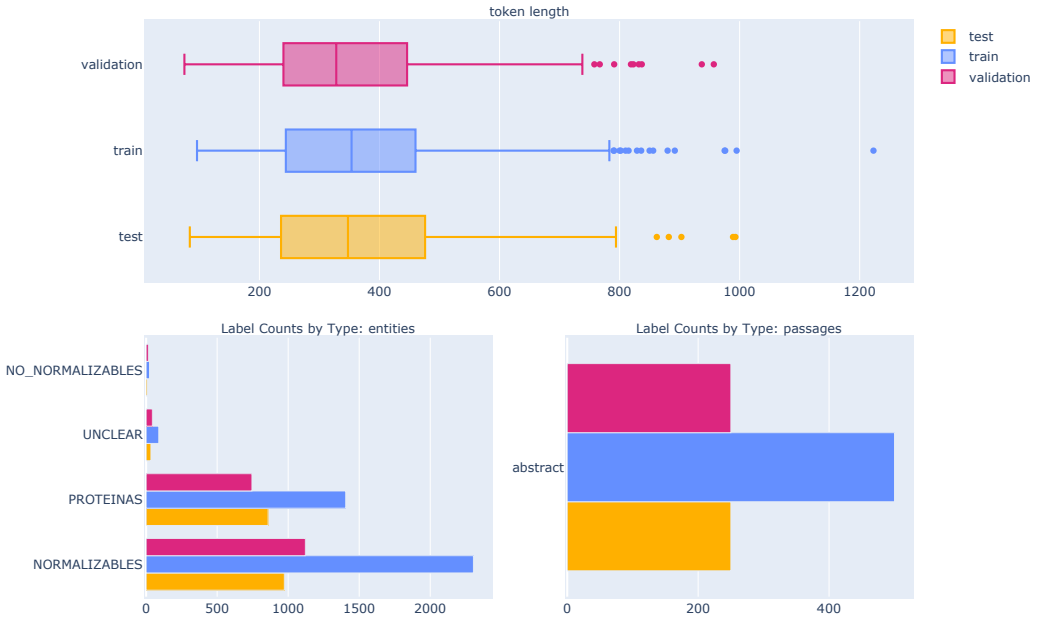

Figure 123: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: PharmaCoNER: Pharmacological Substances, Compounds and Proteins Named Entity Recognition track. This dataset is designed for the PharmaCoNER task, sponsored by Plan de Impulso de las Tecnologías del Lenguaje.

It is a manually classified collection of clinical case studies derived from the Spanish Clinical Case Corpus (SPACCC), an open access electronic library that gathers Spanish medical publications from SciELO (Scientific Electronic Library Online).The annotation of the entire set of entity mentions was carried out by medicinal chemistry experts and it includes the following 4 entity types: NORMALIZABLES, NO_NORMALIZABLES, PROTEINAS and UNCLEAR.

The PharmaCoNER corpus contains a total of 396,988 words and 1,000 clinical cases that have been randomly sampled into 3 subsets. The training set contains 500 clinical cases, while the development and test sets contain 250 clinical cases each.

For further information, please visit https://temu.bsc.es/pharmaconer/ or send an email to encargo-pln-life@bsc.esSUBTRACK 1: NER offset and entity type classification. The first subtrack consists in the classical entity-based or instanced-based evaluation that requires that system outputs match exactly the beginning and end locations of each entity tag, as well as match the entity annotation type of the gold standard annotations.

Homepage: `https://temu.bsc.es/pharmaconer/index.php/datasets/`

URL: `https://temu.bsc.es/pharmaconer/index.php/datasets/`

Licensing: CC_BY_4p0

Languages: Spanish

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

PharmaCoNER Data Card

Figure 124: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: PharmaCoNER: Pharmacological Substances, Compounds and Proteins Named Entity Recognition track. This dataset is designed for the PharmaCoNER task, sponsored by Plan de Impulso de las Tecnologías del Lenguaje.

It is a manually classified collection of clinical case studies derived from the Spanish Clinical Case Corpus (SPACCC), an open access electronic library that gathers Spanish medical publications from SciELO (Scientific Electronic Library Online).The annotation of the entire set of entity mentions was carried out by medicinal chemistry experts and it includes the following 4 entity types: NORMALIZABLES, NO_NORMALIZABLES, PROTEINAS and UNCLEAR.

The PharmaCoNER corpus contains a total of 396,988 words and 1,000 clinical cases that have been randomly sampled into 3 subsets. The training set contains 500 clinical cases, while the development and test sets contain 250 clinical cases each.

For further information, please visit https://temu.bsc.es/pharmaconer/ or send an email to encargo-pln-life@bsc.esSUBTRACK 1: NER offset and entity type classification. The first subtrack consists in the classical entity-based or instanced-based evaluation that requires that system outputs match exactly the beginning and end locations of each entity tag, as well as match the entity annotation type of the gold standard annotations.

Homepage: `https://temu.bsc.es/pharmaconer/index.php/datasets/`

URL: `https://temu.bsc.es/pharmaconer/index.php/datasets/`

Licensing: CC_BY_4p0

Languages: Spanish

Tasks: text classification

Schemas: `TEXT KB`

Splits: train, test, validation

PhoNER_COVID19 Data Card

Figure 125: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description PhoNER_COVID19 is a dataset for recognizing COVID-19 related named entities in Vietnamese, consisting of 35K entities over 10K sentences. We define 10 entity types with the aim of extracting key information related to COVID-19 patients, which are especially useful in downstream applications. In general, these entity types can be used in the context of not only the COVID-19 pandemic but also in other future epidemics

Homepage: `https://github.com/VinAIResearch/PhoNER_COVID19`

URL: `https://github.com/VinAIResearch/PhoNER_COVID19`

Licensing: Custom license

Languages: Vietnamese

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

PMC-Patients Data Card

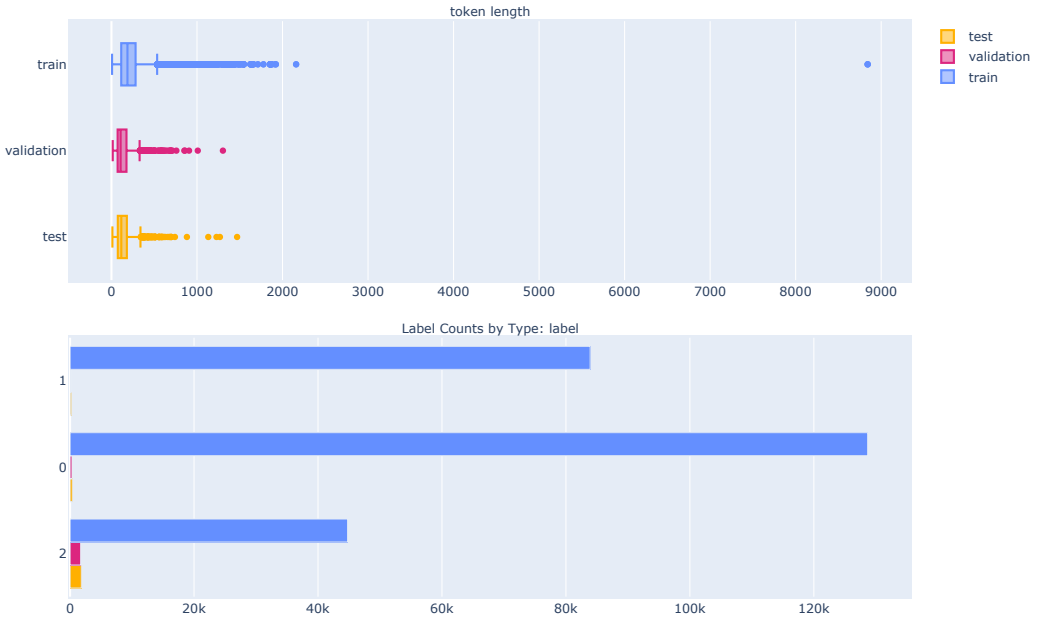

Figure 126: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: This dataset is used for calculating the similarity between two patient descriptions.

Homepage: `https://github.com/zhao-zy15/PMC-Patients`

URL: `https://github.com/zhao-zy15/PMC-Patients`

Licensing: CC_BY_NC_SA_4p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train, test, validation

ProGene Data Card

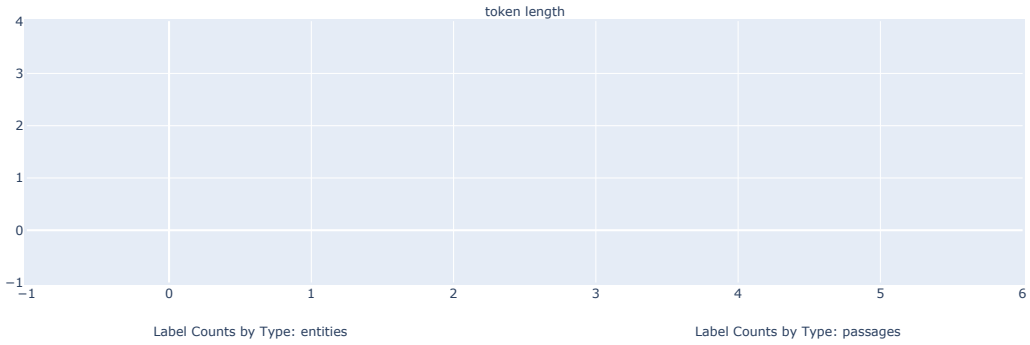

Figure 127: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The Protein/Gene corpus was developed at the JULIE Lab Jena under supervision of Prof. Udo Hahn.The executing scientist was Dr. Joachim Wermter.The main annotator was Dr. Rico Pusch who is an expert in biology.The corpus was developed in the context of the StemNet project (http://www.stemnet.de/).

Homepage: `https://zenodo.org/record/3698568#.YlVHqdNBxeg`

URL: `https://zenodo.org/record/3698568#.YlVHqdNBxeg`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: validation, train, test

PUBHEALTH Data Card

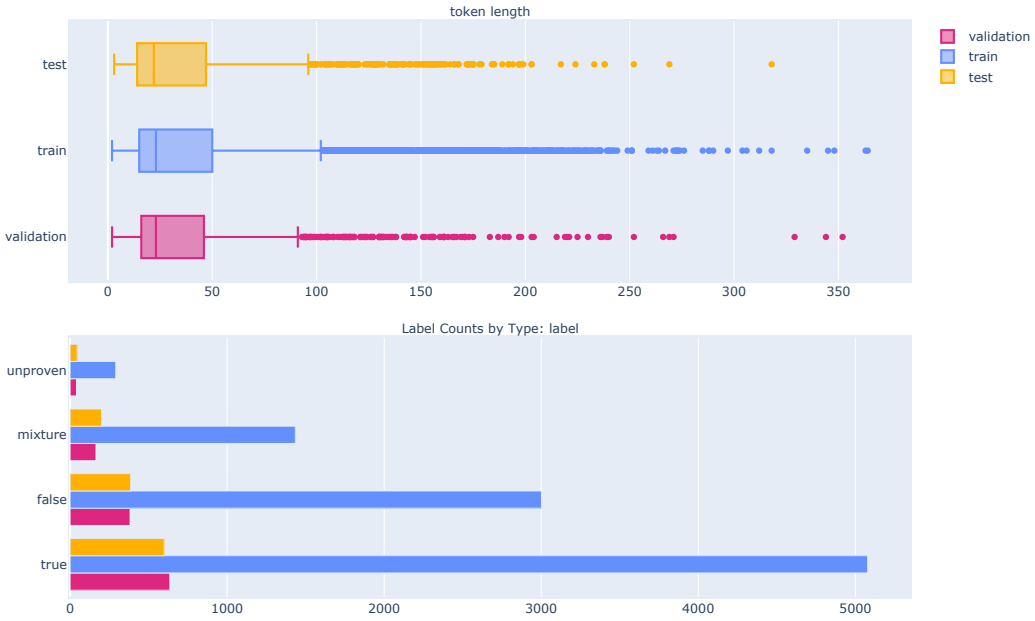

Figure 128: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: A dataset of 11,832 claims for fact-checking, which are related a range of health topics including biomedical subjects (e.g., infectious diseases, stem cell research), government healthcare policy(e.g., abortion, mental health, women's health), and other public health-related stories

Homepage: `https://github.com/neemakot/Health-Fact-Checking/tree/master/data`

URL: `https://github.com/neemakot/Health-Fact-Checking/tree/master/data`

Licensing: MIT

Languages: English

Tasks:

Schemas: `PAIRS`

Splits: train, test, validation

PubMedQA (Artificial) Data Card

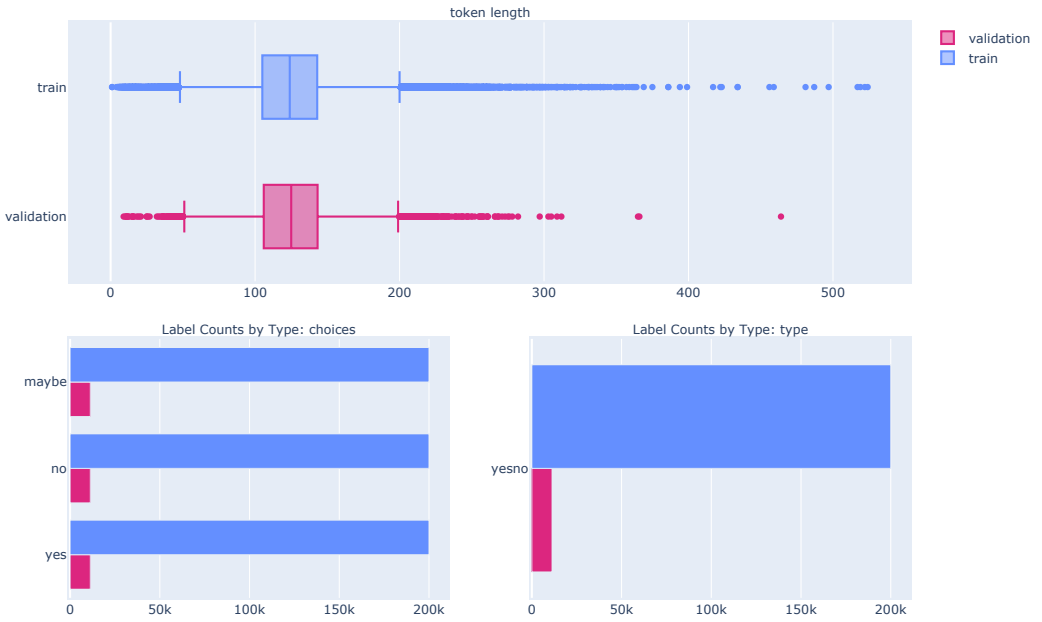

Figure 129: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: PubMedQA is a novel biomedical question answering (QA) dataset collected from PubMed abstracts.The task of PubMedQA is to answer research biomedical questions with yes/no/maybe using the corresponding abstracts.PubMedQA has 1k expert-annotated (PQA-L), 61.2k unlabeled (PQA-U) and 211.3k artificially generated QA instances (PQA-A).

Each PubMedQA instance is composed of: (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding PubMed abstract without its conclusion, (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion.

PubMedQA is the first QA dataset where reasoning over biomedical research texts,especially their quantitative contents, is required to answer the questions.PubMedQA datasets comprise of 3 different subsets: (1) PubMedQA Labeled (PQA-L): A labeled PubMedQA subset comprises of 1k manually annotated yes/no/maybe QA data collected from PubMed articles. (2) PubMedQA Artificial (PQA-A): An artificially labelled PubMedQA subset comprises of 211.3k PubMed articles with automatically generated questions from the statement titles and yes/no answer labels generated using a simple heuristic. (3) PubMedQA Unlabeled (PQA-U): An unlabeled PubMedQA subset comprises of 61.2k context-question pairs data collected from PubMed articles.

Homepage: `https://github.com/pubmedqa/pubmedqa`

URL: `https://github.com/pubmedqa/pubmedqa`

Licensing: MIT

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, validation

PubMedQA (Fold0) Data Card

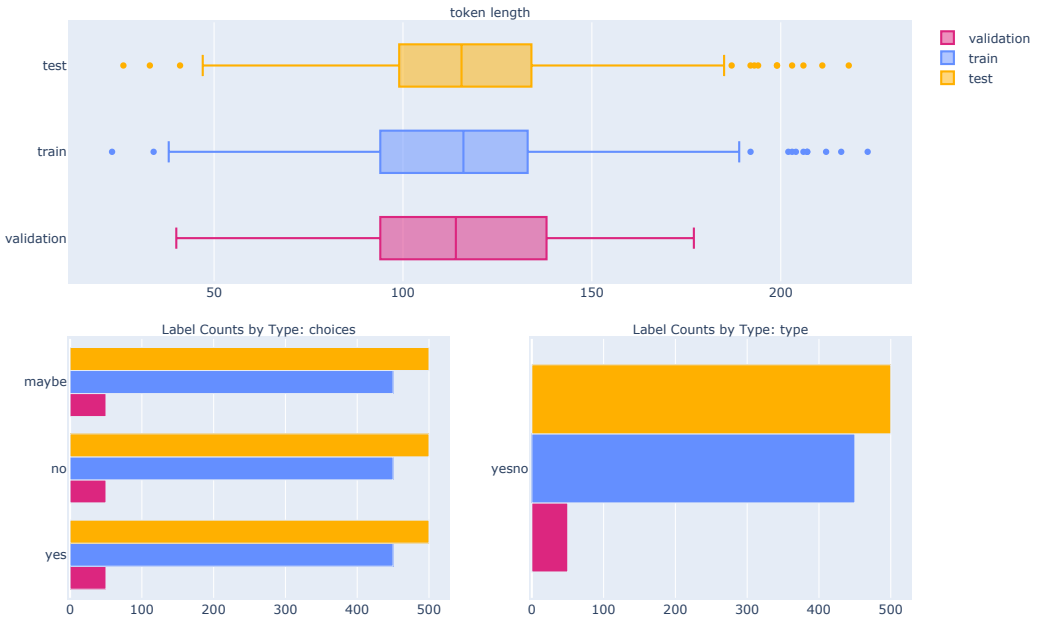

Figure 130: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: PubMedQA is a novel biomedical question answering (QA) dataset collected from PubMed abstracts.The task of PubMedQA is to answer research biomedical questions with yes/no/maybe using the corresponding abstracts.PubMedQA has 1k expert-annotated (PQA-L), 61.2k unlabeled (PQA-U) and 211.3k artificially generated QA instances (PQA-A).

Each PubMedQA instance is composed of: (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding PubMed abstract without its conclusion, (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion.

PubMedQA is the first QA dataset where reasoning over biomedical research texts,especially their quantitative contents, is required to answer the questions.PubMedQA datasets comprise of 3 different subsets: (1) PubMedQA Labeled (PQA-L): A labeled PubMedQA subset comprises of 1k manually annotated yes/no/maybe QA data collected from PubMed articles. (2) PubMedQA Artificial (PQA-A): An artificially labelled PubMedQA subset comprises of 211.3k PubMed articles with automatically generated questions from the statement titles and yes/no answer labels generated using a simple heuristic. (3) PubMedQA Unlabeled (PQA-U): An unlabeled PubMedQA subset comprises of 61.2k context-question pairs data collected from PubMed articles.

Homepage: `https://github.com/pubmedqa/pubmedqa`

URL: `https://github.com/pubmedqa/pubmedqa`

Licensing: MIT

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train, validation, test

PubMedQA (Unlabeled) Data Card

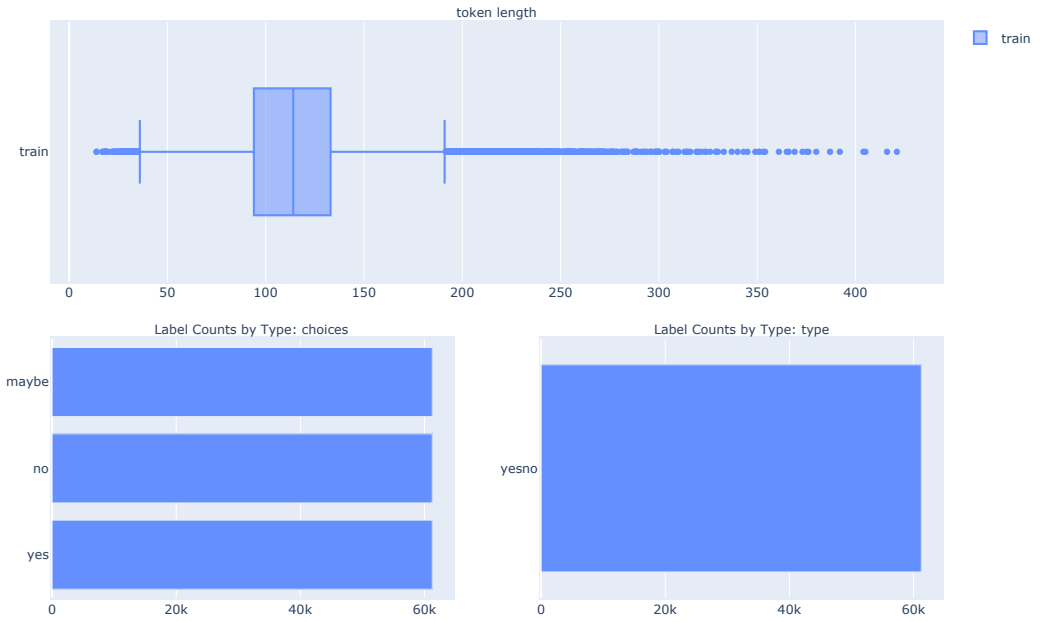

Figure 131: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: PubMedQA is a novel biomedical question answering (QA) dataset collected from PubMed abstracts.The task of PubMedQA is to answer research biomedical questions with yes/no/maybe using the corresponding abstracts.PubMedQA has 1k expert-annotated (PQA-L), 61.2k unlabeled (PQA-U) and 211.3k artificially generated QA instances (PQA-A).

Each PubMedQA instance is composed of: (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding PubMed abstract without its conclusion, (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion.

PubMedQA is the first QA dataset where reasoning over biomedical research texts,especially their quantitative contents, is required to answer the questions.PubMedQA datasets comprise of 3 different subsets: (1) PubMedQA Labeled (PQA-L): A labeled PubMedQA subset comprises of 1k manually annotated yes/no/maybe QA data collected from PubMed articles. (2) PubMedQA Artificial (PQA-A): An artificially labelled PubMedQA subset comprises of 211.3k PubMed articles with automatically generated questions from the statement titles and yes/no answer labels generated using a simple heuristic. (3) PubMedQA Unlabeled (PQA-U): An unlabeled PubMedQA subset comprises of 61.2k context-question pairs data collected from PubMed articles.

Homepage: `https://github.com/pubmedqa/pubmedqa`

URL: `https://github.com/pubmedqa/pubmedqa`

Licensing: MIT

Languages: English

Tasks: question answering

Schemas: `QA`

Splits: train

QUAERO (EMEA) Data Card

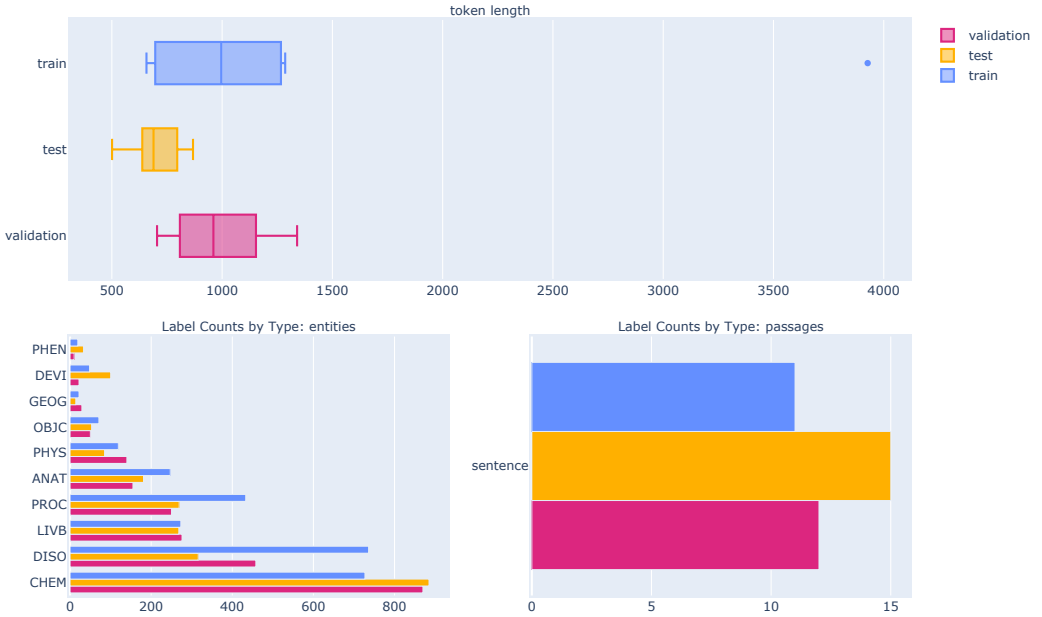

Figure 132: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The QUAERO French Medical Corpus has been initially developed as a resource for named entity recognition and normalization [1]. It was then improved with the purpose of creating a gold standard set of normalized entities for French biomedical text, that was used in the CLEF eHealth evaluation lab [2][3].

A selection of MEDLINE titles and EMEA documents were manually annotated. The annotation process was guided by concepts in the Unified Medical Language System (UMLS):1. Ten types of clinical entities, as defined by the following UMLS Semantic Groups (Bodenreider and McCray 2003) were annotated: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures.2. The annotations were made in a comprehensive fashion, so that nested entities were marked, and entities could be mapped to more than one UMLS concept. In particular: (a) If a mention can refer to more than one Semantic Group, all the relevant Semantic Groups should be annotated. For instance, the mention “récidive” (recurrence) in the phrase “prévention des récidives” (recurrence prevention) should be annotated with the category “DISORDER” (CUI C2825055) and the category “PHENOMENON” (CUI C0034897); (b) If a mention can refer to more than one UMLS concept within the same Semantic Group, all the relevant concepts should be annotated. For instance, the mention “maniaques” (obsessive) in the phrase “patients maniaques” (obsessive patients) should be annotated with CUIs C0564408 and C0338831 (category “DISORDER”); (c) Entities which span overlaps with that of another entity should still be annotated. For instance, in the phrase “infarctus du myocarde” (myocardial infarction), the mention “myocarde” (myocardium) should be annotated with category “ANATOMY” (CUI C0027061) and the mention “infarctus du myocarde” should be annotated with category “DISORDER” (CUI C0027051)

The QUAERO French Medical Corpus BioC release comprises a subset of the QUAERO French Medical corpus, as follows:Training data (BRAT version used in CLEF eHealth 2015 task 1b as training data): - MEDLINE_train_bioc file: 833 MEDLINE titles, annotated with normalized entities in the BioC format - EMEA_train_bioc file: 3 EMEA documents, segmented into 11 sub-documents, annotated with normalized entities in the BioC format Development data (BRAT version used in CLEF eHealth 2015 task 1b as test data and in CLEF eHealth 2016 task 2 as development data): - MEDLINE_dev_bioc file: 832 MEDLINE titles, annotated with normalized entities in the BioC format- EMEA_dev_bioc file: 3 EMEA documents, segmented into 12 sub-documents, annotated with normalized entities in the BioC format Test data (BRAT version used in CLEF eHealth 2016 task 2 as

test data): - MEDLINE_test_bioc folder: 833 MEDLINE titles, annotated with normalized entities in the BioC format - EMEA folder_test_bioc: 4 EMEA documents, segmented into 15 sub-documents, annotated with normalized entities in the BioC format This release of the QUAERO French medical corpus, BioC version, comes in the BioC format, through automatic conversion from the original BRAT format obtained with the Brat2BioC tool https://bitbucket.org/nicta_biomed/brat2bioc developped by Jimeno Yepes et al. Antonio Jimeno Yepes, Mariana Neves, Karin Verspoor Brat2BioC: conversion tool between brat and BioCBioCreative IV track 1 - BioC: The BioCreative Interoperability Initiative, 2013Please note that the original version of the QUAERO corpus distributed in the CLEF eHealth challenge 2015 and 2016 came in the BRAT stand alone format. It was distributed with the CLEF eHealth evaluation tool. This original distribution of the QUAERO French Medical corpus is available separately from https://quaerofrenchmed.limsi.fr

All questions regarding the task or data should be addressed to aurelie.neveol@limsi.fr

Homepage: `https://quaerofrenchmed.limsi.fr/`

URL: `https://quaerofrenchmed.limsi.fr/`

Licensing: GFDL_1p3

Languages: French

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

QUAERO (Medline) Data Card

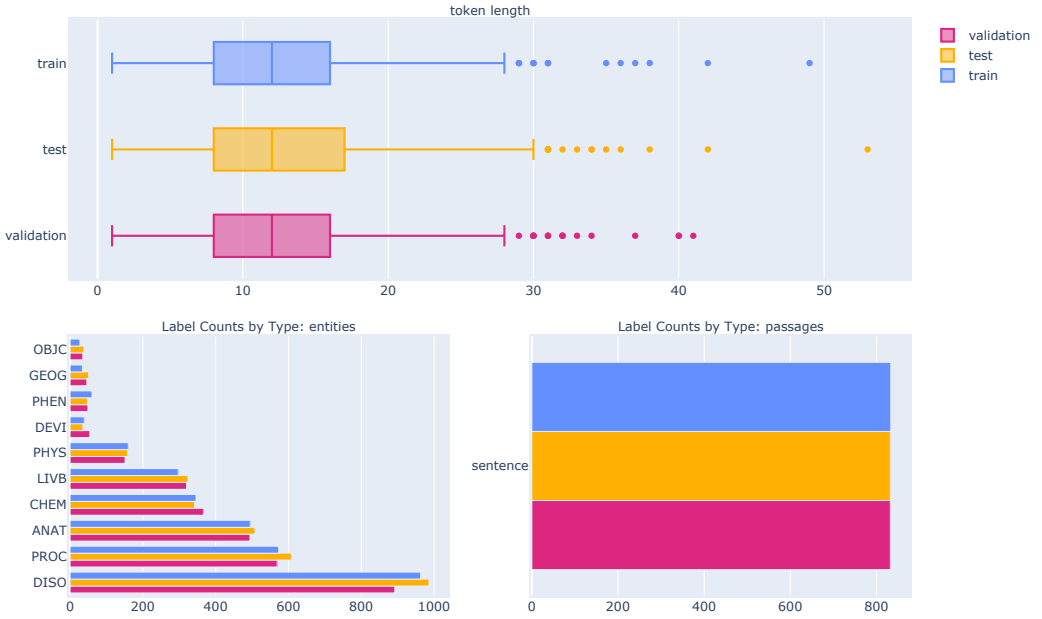

Figure 133: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The QUAERO French Medical Corpus has been initially developed as a resource for named entity recognition and normalization [1]. It was then improved with the purpose of creating a gold standard set of normalized entities for French biomedical text, that was used in the CLEF eHealth evaluation lab [2][3].

A selection of MEDLINE titles and EMEA documents were manually annotated. The annotation process was guided by concepts in the Unified Medical Language System (UMLS):1. Ten types of clinical entities, as defined by the following UMLS Semantic Groups (Bodenreider and McCray 2003) were annotated: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures.2. The annotations were made in a comprehensive fashion, so that nested entities were marked, and entities could be mapped to more than one UMLS concept. In particular: (a) If a mention can refer to more than one Semantic Group, all the relevant Semantic Groups should be annotated. For instance, the mention “récidive” (recurrence) in the phrase “prévention des récidives” (recurrence prevention) should be annotated with the category “DISORDER” (CUI C2825055) and the category “PHENOMENON” (CUI C0034897); (b) If a mention can refer to more than one UMLS concept within the same Semantic Group, all the relevant concepts should be annotated. For instance, the mention “maniaques” (obsessive) in the phrase “patients maniaques” (obsessive patients) should be annotated with CUIs C0564408 and C0338831 (category “DISORDER”); (c) Entities which span overlaps with that of another entity should still be annotated. For instance, in the phrase “infarctus du myocarde” (myocardial infarction), the mention “myocarde” (myocardium) should be annotated with category “ANATOMY” (CUI C0027061) and the mention “infarctus du myocarde” should be annotated with category “DISORDER” (CUI C0027051)

The QUAERO French Medical Corpus BioC release comprises a subset of the QUAERO French Medical corpus, as follows:Training data (BRAT version used in CLEF eHealth 2015 task 1b as training data): - MEDLINE_train_bioc file: 833 MEDLINE titles, annotated with normalized entities in the BioC format - EMEA_train_bioc file: 3 EMEA documents, segmented into 11 sub-documents, annotated with normalized entities in the BioC format Development data (BRAT version used in CLEF eHealth 2015 task 1b as test data and in CLEF eHealth 2016 task 2 as development data): - MEDLINE_dev_bioc file: 832 MEDLINE titles, annotated with normalized entities in the BioC format- EMEA_dev_bioc file: 3 EMEA documents, segmented into 12 sub-documents, annotated with normalized entities in the BioC format Test data (BRAT version used in CLEF eHealth 2016 task 2 as

test data): - MEDLINE_test_bioc folder: 833 MEDLINE titles, annotated with normalized entities in the BioC format - EMEA folder_test_bioc: 4 EMEA documents, segmented into 15 sub-documents, annotated with normalized entities in the BioC format This release of the QUAERO French medical corpus, BioC version, comes in the BioC format, through automatic conversion from the original BRAT format obtained with the Brat2BioC tool https://bitbucket.org/nicta_biomed/brat2bioc developped by Jimeno Yepes et al. Antonio Jimeno Yepes, Mariana Neves, Karin Verspoor Brat2BioC: conversion tool between brat and BioCBioCreative IV track 1 - BioC: The BioCreative Interoperability Initiative, 2013Please note that the original version of the QUAERO corpus distributed in the CLEF eHealth challenge 2015 and 2016 came in the BRAT stand alone format. It was distributed with the CLEF eHealth evaluation tool. This original distribution of the QUAERO French Medical corpus is available separately from https://quaerofrenchmed.limsi.fr

All questions regarding the task or data should be addressed to aurelie.neveol@limsi.fr

Homepage: `https://quaerofrenchmed.limsi.fr/`

URL: `https://quaerofrenchmed.limsi.fr/`

Licensing: GFDL_1p3

Languages: French

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test, validation

SCAI Chemical Data Card

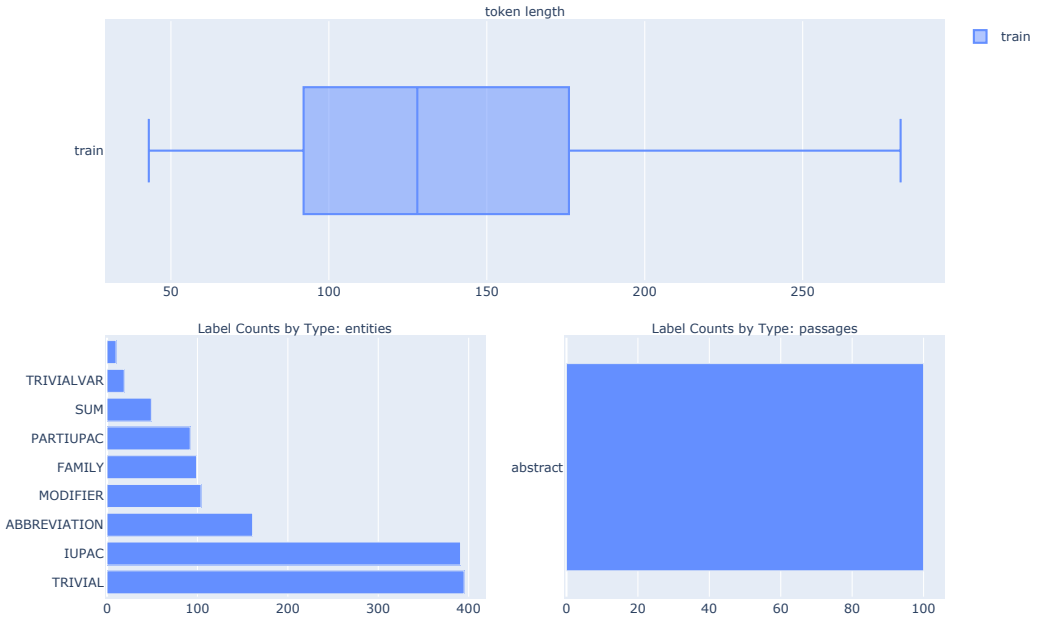

Figure 134: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: SCAI Chemical is a corpus of MEDLINE abstracts that has been annotated to give an overview of the different chemical name classes found in MEDLINE text.

Homepage: `https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/corpora-for-chemical-entity-recognition.html`

URL: `https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/corpora-for-chemical-entity-recognition.html`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

SCAI Disease Data Card

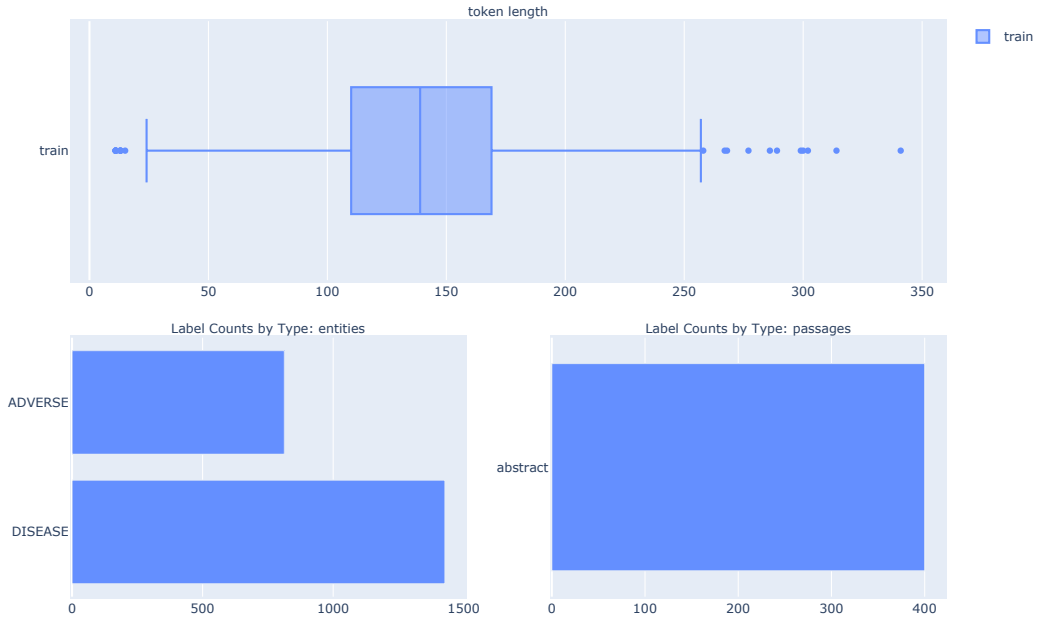

Figure 135: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: SCAI Disease is a dataset annotated in 2010 with mentions of diseases and adverse effects. It is a corpus containing 400 randomly selected MEDLINE abstracts generated using 'Disease OR Adverse effect' as a PubMed query. This evaluation corpus was annotated by two individuals who hold a Master's degree in life sciences.

Homepage: `https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/corpus-for-disease-names-and-adverse-effects.html`

URL: `https://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/corpus-for-disease-names-and-adverse-effects.html`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train

SciCite Data Card

Figure 136: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: SciCite is a dataset of 11K manually annotated citation intents based on citation context in the computer science and biomedical domains.

Homepage: `https://allenai.org/data/scicite`

URL: `https://allenai.org/data/scicite`

Licensing: UNKNOWN

Languages: English

Tasks: text classification

Schemas: `TEXT`

Splits: train, test, validation

SciELO (Spanish) Data Card

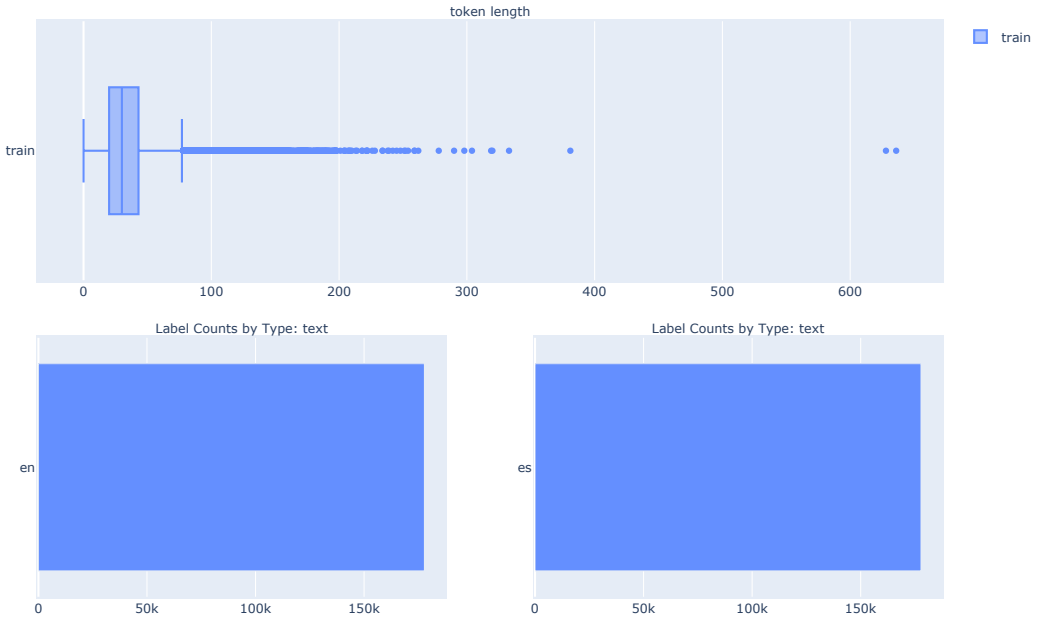

Figure 137: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: A parallel corpus of full-text scientific articles collected from Scielo database in the following languages: English, Portuguese and Spanish. The corpus is sentence aligned for all language pairs, as well as trilingual aligned for a small subset of sentences. Alignment was carried out using the Hunalign algorithm.

Homepage: `https://sites.google.com/view/felipe-soares/datasets#h.p_92uSCyAjWSRB`

URL: `https://sites.google.com/view/felipe-soares/datasets#h.p_92uSCyAjWSRB`

Licensing: CC_BY_4p0

Languages: English, Spanish

Tasks: translation

Schemas: `T2T`

Splits: train

SciELO (Portuguese) Data Card

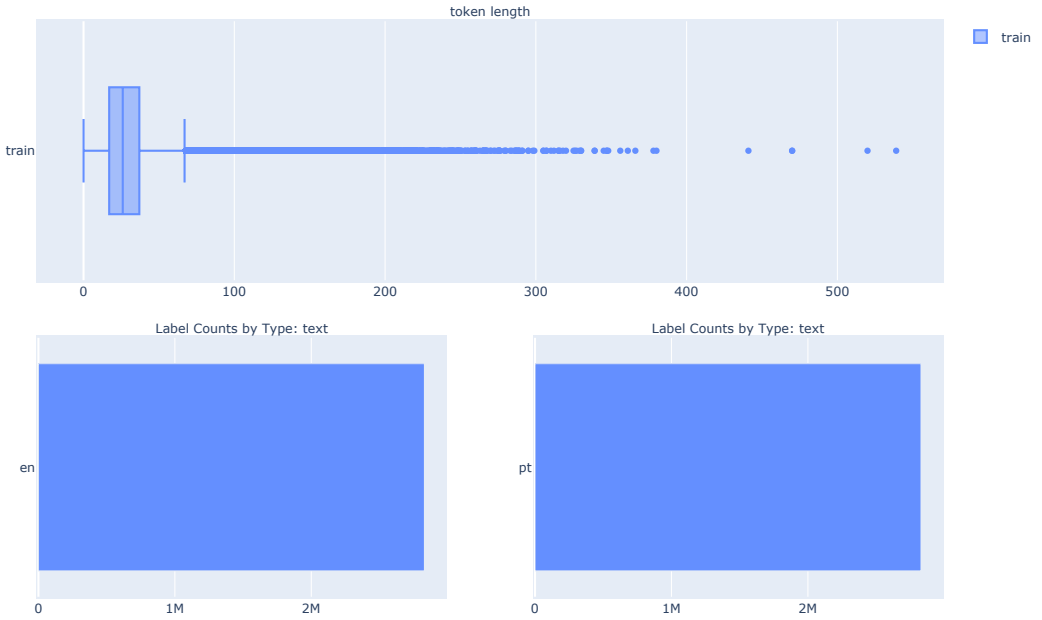

Figure 138: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description A parallel corpus of full-text scientific articles collected from Scielo database in the following languages: English, Portuguese and Spanish. The corpus is sentence aligned for all language pairs, as well as trilingual aligned for a small subset of sentences. Alignment was carried out using the Hunalign algorithm.

Homepage: `https://sites.google.com/view/felipe-soares/datasets#h.p_92uSCyAjWSRB`

URL: `https://sites.google.com/view/felipe-soares/datasets#h.p_92uSCyAjWSRB`

Licensing: CC_BY_4p0

Languages: English, Portuguese

Tasks: translation

Schemas: `T2T`

Splits: train

SciFact (Label Prediction) Data Card

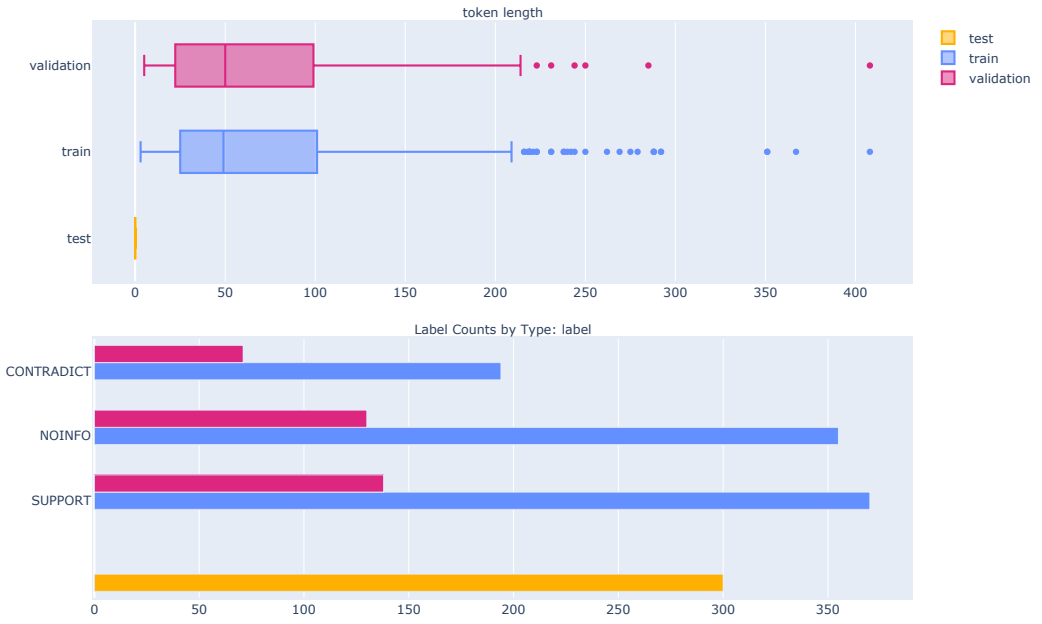

Figure 139: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: SciFact is a dataset of 1.4K expert-written scientific claims paired with evidence-containing abstracts, and annotated with labels and rationales. This config has abstracts and document ids.

Homepage: `https://scifact.apps.allenai.org/`

URL: `https://scifact.apps.allenai.org/`

Licensing: CC_BY_NC_2p0

Languages: English

Tasks: textual entailment

Schemas: `TE`

Splits: train, test, validation

SciFact (Rationale) Data Card

Figure 140: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: SciFact is a dataset of 1.4K expert-written scientific claims paired with evidence-containing abstracts, and annotated with labels and rationales. This config has abstracts and document ids.

Homepage: `https://scifact.apps.allenai.org/`

URL: `https://scifact.apps.allenai.org/`

Licensing: CC_BY_NC_2p0

Languages: English

Tasks: textual entailment

Schemas: `TE`

Splits: train, test, validation

SciTail Data Card

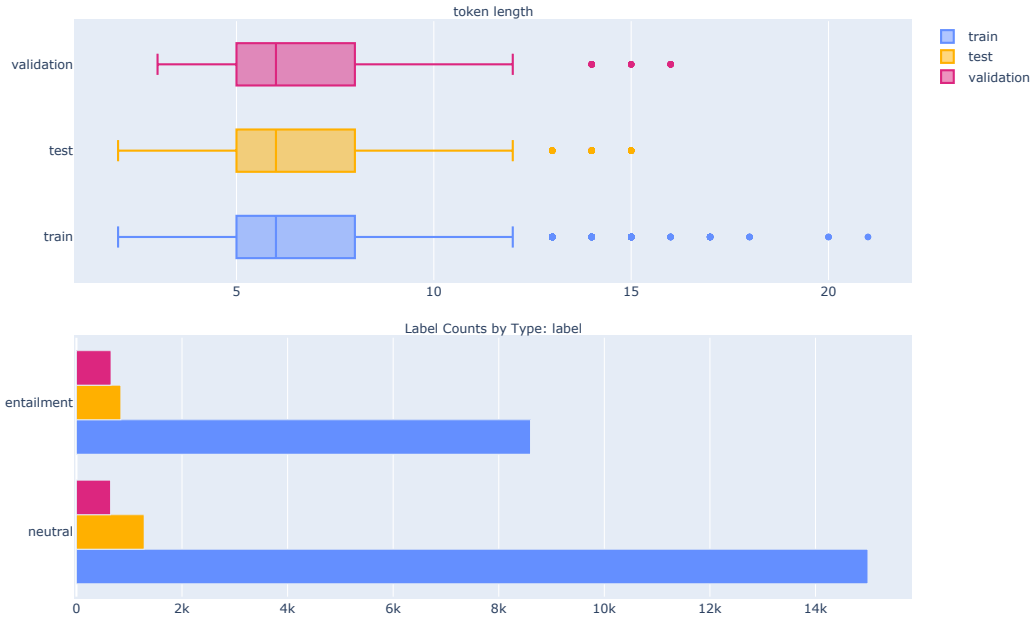

Figure 141: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The SciTail dataset is an entailment dataset created from multiple-choice science exams and web sentences. Each question and the correct answer choice are converted into an assertive statement to form the hypothesis. We use information retrieval to obtain relevant text from a large text corpus of web sentences, and use these sentences as a premise P. We crowdsource the annotation of such premise-hypothesis pair as supports (entails) or not (neutral), in order to create the SciTail dataset. The dataset contains 27,026 examples with 10,101 examples with entails label and 16,925 examples with neutral label.

Homepage: `https://allenai.org/data/scitail`

URL: `https://allenai.org/data/scitail`

Licensing: APACHE_2p0

Languages: English

Tasks: textual entailment

Schemas: `TE`

Splits: train, test, validation

SETH Corpus Data Card

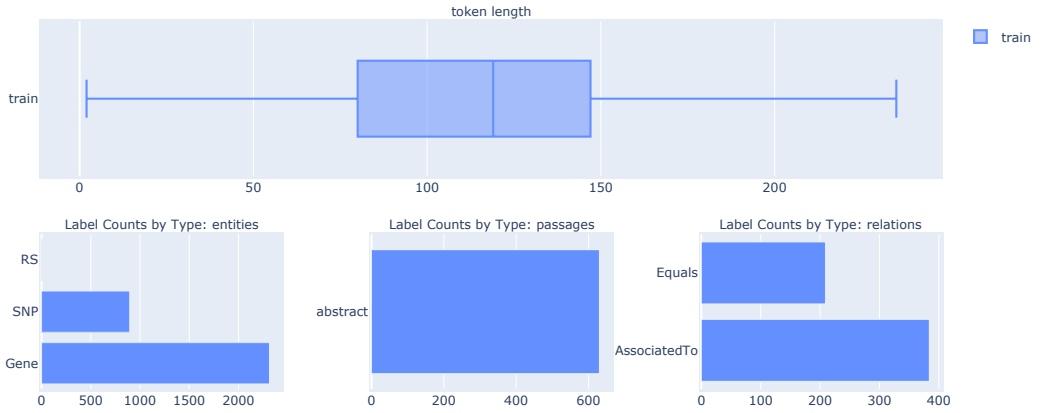

Figure 142: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: SNP named entity recognition corpus consisting of 630 PubMed citations.

Homepage: `https://github.com/rockt/SETH`

URL: `https://github.com/rockt/SETH`

Licensing: APACHE_2p0

Languages: English

Tasks: relation extraction, named entity recognition

Schemas: `KB`

Splits: train

SPL ADR (Train) Data Card

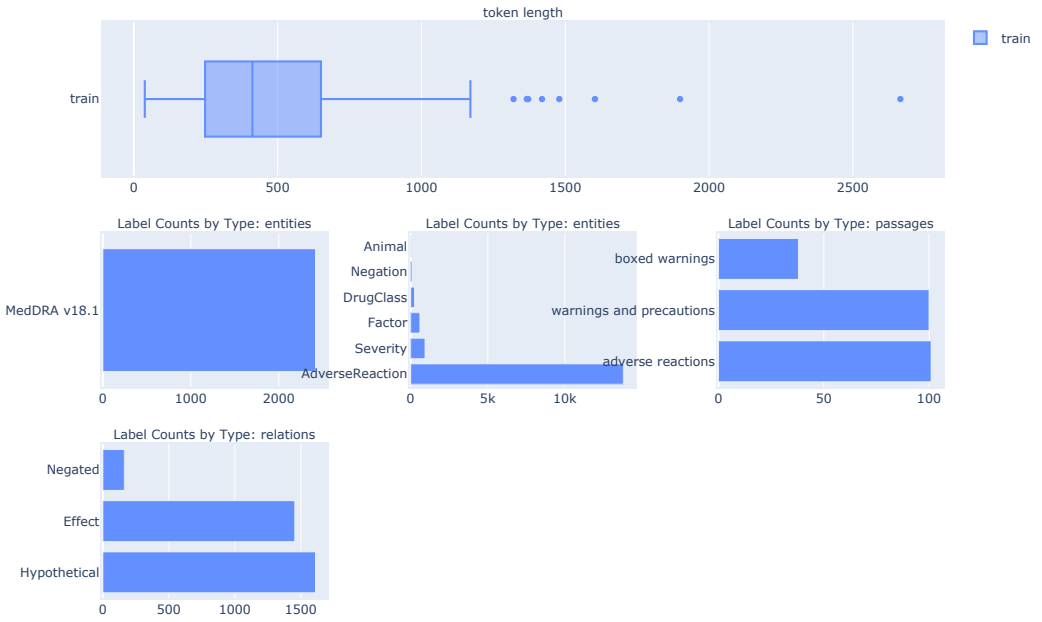

Figure 143: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The United States Food and Drug Administration (FDA) partnered with the National Library of Medicine to create a pilot dataset containing standardised information about known adverse reactions for 200 FDA-approved drugs. The Structured Product Labels (SPLs),the documents FDA uses to exchange information about drugs and other products, were manually annotated for adverse reactions at the mention level to facilitate development and evaluation of text mining tools for extraction of ADRs from all SPLs. The ADRs were then normalised to the Unified Medical Language System (UMLS) and to the Medical Dictionary for Regulatory Activities (MedDRA).

Homepage: `https://bionlp.nlm.nih.gov/tac2017adversereactions/`

URL: `https://bionlp.nlm.nih.gov/tac2017adversereactions/`

Licensing: CC0_1p0

Languages: English

Tasks: relation extraction, named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

SPL ADR (Unannotated) Data Card

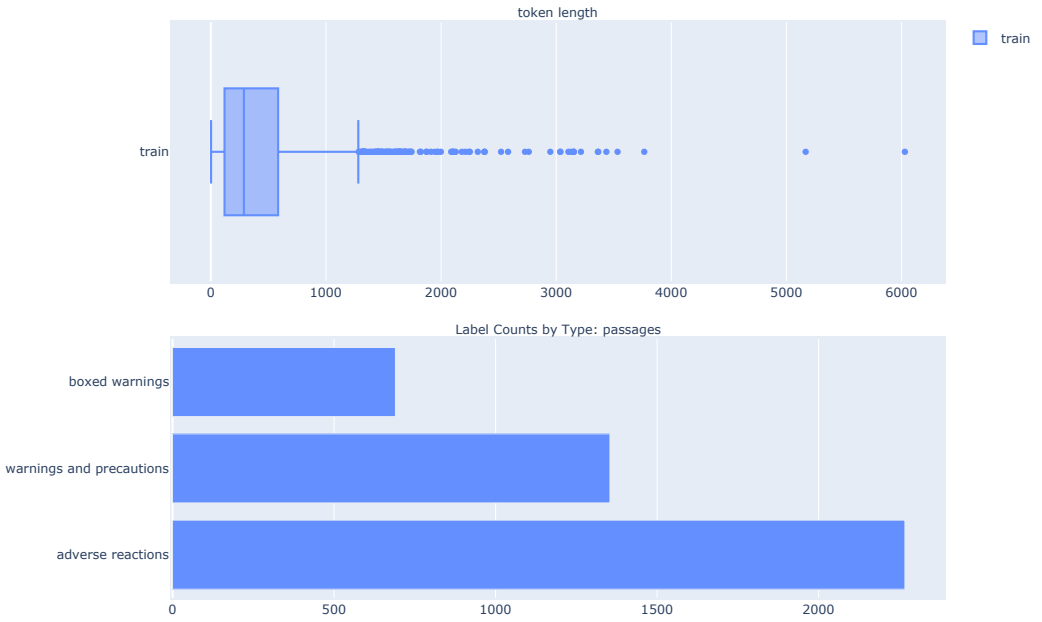

Figure 144: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The United States Food and Drug Administration (FDA) partnered with the National Library of Medicine to create a pilot dataset containing standardised information about known adverse reactions for 200 FDA-approved drugs. The Structured Product Labels (SPLs),the documents FDA uses to exchange information about drugs and other products, were manually annotated for adverse reactions at the mention level to facilitate development and evaluation of text mining tools for extraction of ADRs from all SPLs. The ADRs were then normalised to the Unified Medical Language System (UMLS) and to the Medical Dictionary for Regulatory Activities (MedDRA).

Homepage: `https://bionlp.nlm.nih.gov/tac2017adversereactions/`

URL: `https://bionlp.nlm.nih.gov/tac2017adversereactions/`

Licensing: CC0_1p0

Languages: English

Tasks: relation extraction, named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

Swedish Medical NER (1177 Vårdguiden) Data Card

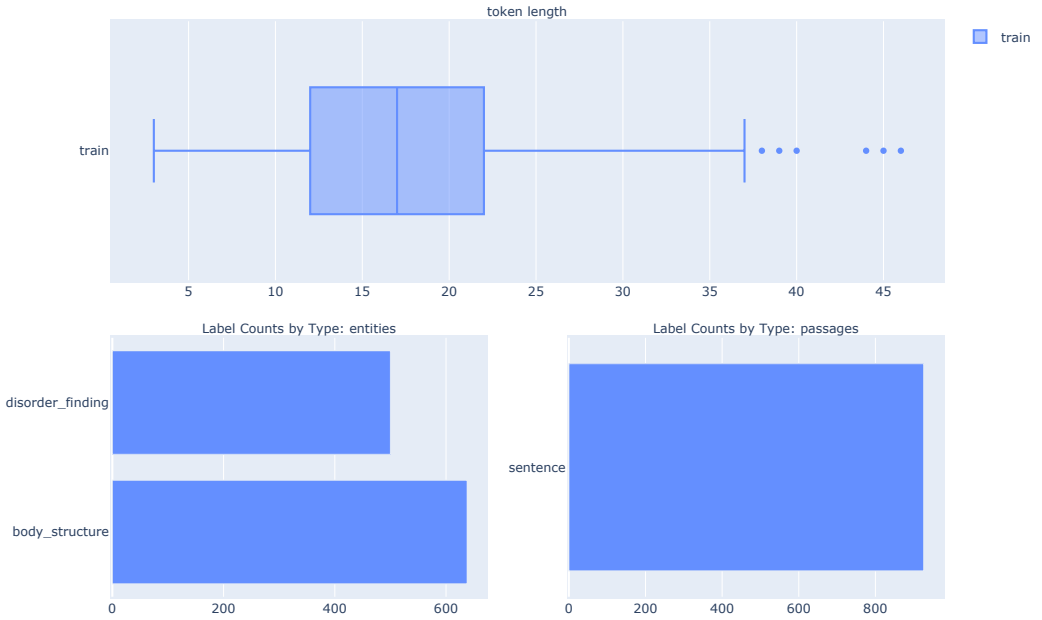

Figure 145: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Named Entity Recognition dataset on medical text in Swedish. It consists three subsets which are in turn derived from three different sources respectively: the Swedish Wikipedia (a.k.a. wiki), Läkartidningen (a.k.a. lt), and 1177 Vårdguiden (a.k.a. 1177). While the Swedish Wikipedia and Läkartidningen subsets in total contains over 790000 sequences with 60 characters each, the 1177 Vårdguiden subset is manually annotated and contains 927 sentences, 2740 annotations, out of which 1574 are disorder and findings, 546 are pharmaceutical drug, and 620 are body structure.Texts from both Swedish Wikipedia and Läkartidningen were automatically annotated using a list of medical seed terms. Sentences from 1177 Vårdguiden were manually annotated.

Homepage: `https://github.com/olofmogren/biomedical-ner-data-swedish/`

URL: `https://github.com/olofmogren/biomedical-ner-data-swedish/`

Licensing: CC_BY_SA_4p0

Languages: Swedish

Tasks: named entity recognition

Schemas: `KB`

Splits: train

Swedish Medical NER (Läkartidningen) Data Card

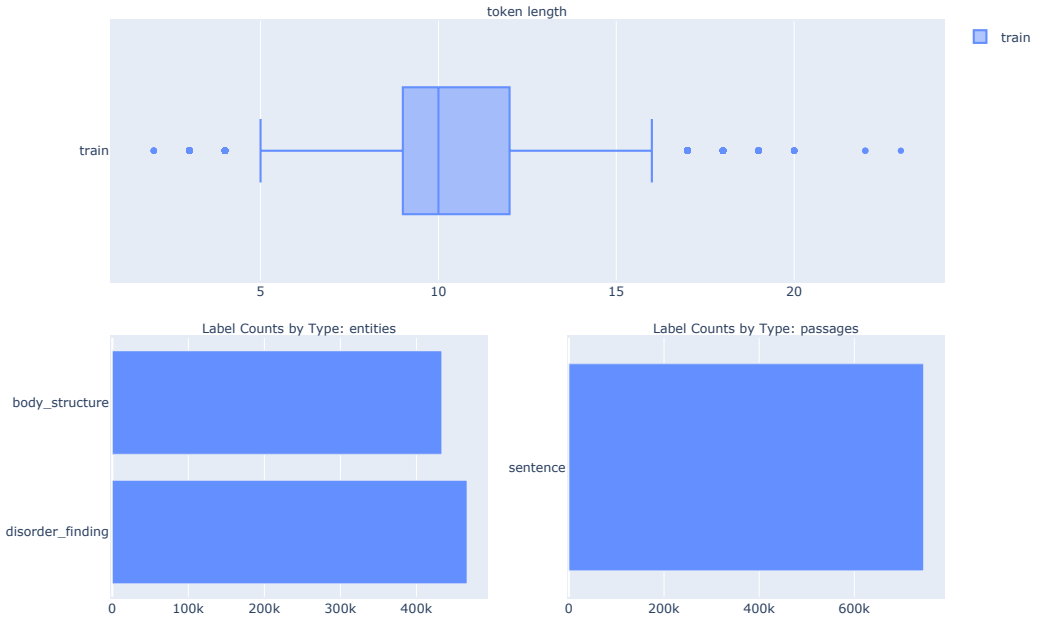

Figure 146: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Named Entity Recognition dataset on medical text in Swedish. It consists three subsets which are in turn derived from three different sources respectively: the Swedish Wikipedia (a.k.a. wiki), Läkartidningen (a.k.a. lt), and 1177 Vårdguiden (a.k.a. 1177). While the Swedish Wikipedia and Läkartidningen subsets in total contains over 790000 sequences with 60 characters each, the 1177 Vårdguiden subset is manually annotated and contains 927 sentences, 2740 annotations, out of which 1574 are disorder and findings, 546 are pharmaceutical drug, and 620 are body structure.Texts from both Swedish Wikipedia and Läkartidningen were automatically annotated using a list of medical seed terms. Sentences from 1177 Vårdguiden were manually annotated.

Homepage: `https://github.com/olofmogren/biomedical-ner-data-swedish/`

URL: `https://github.com/olofmogren/biomedical-ner-data-swedish/`

Licensing: CC_BY_SA_4p0

Languages: Swedish

Tasks: named entity recognition

Schemas: `KB`

Splits: train

Swedish Medical NER (Swedish Wikipedia) Data Card

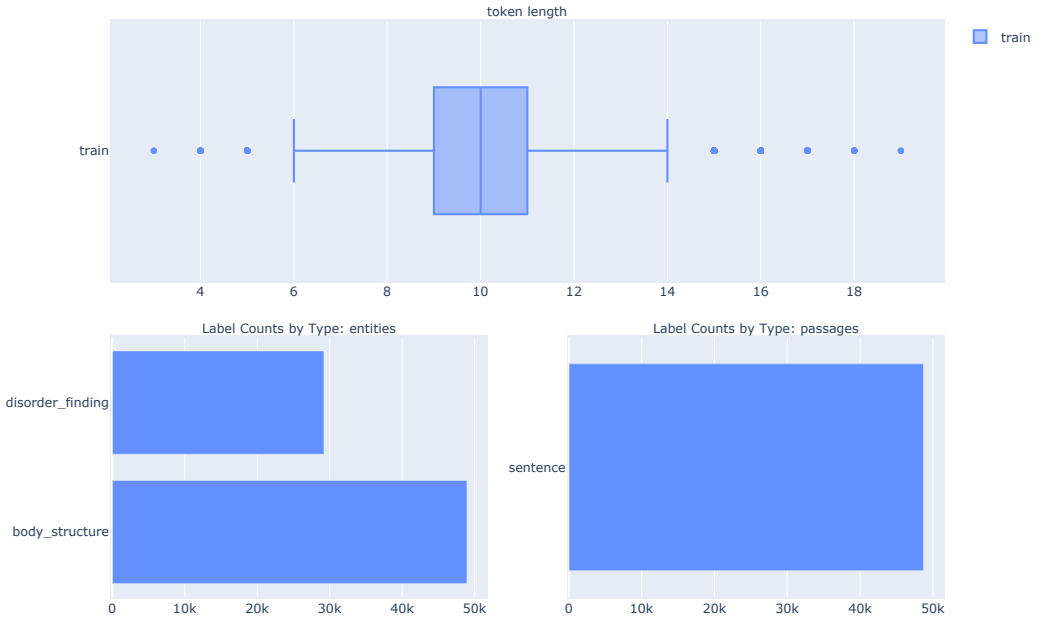

Figure 147: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: Named Entity Recognition dataset on medical text in Swedish. It consists three subsets which are in turn derived from three different sources respectively: the Swedish Wikipedia (a.k.a. wiki), Läkartidningen (a.k.a. lt), and 1177 Vårdguiden (a.k.a. 1177). While the Swedish Wikipedia and Läkartidningen subsets in total contains over 790000 sequences with 60 characters each, the 1177 Vårdguiden subset is manually annotated and contains 927 sentences, 2740 annotations, out of which 1574 are disorder and findings, 546 are pharmaceutical drug, and 620 are body structure.Texts from both Swedish Wikipedia and Läkartidningen were automatically annotated using a list of medical seed terms. Sentences from 1177 Vårdguiden were manually annotated.

Homepage: `https://github.com/olofmogren/biomedical-ner-data-swedish/`

URL: `https://github.com/olofmogren/biomedical-ner-data-swedish/`

Licensing: CC_BY_SA_4p0

Languages: Swedish

Tasks: named entity recognition

Schemas: `KB`

Splits: train

tmVar v1 Data Card

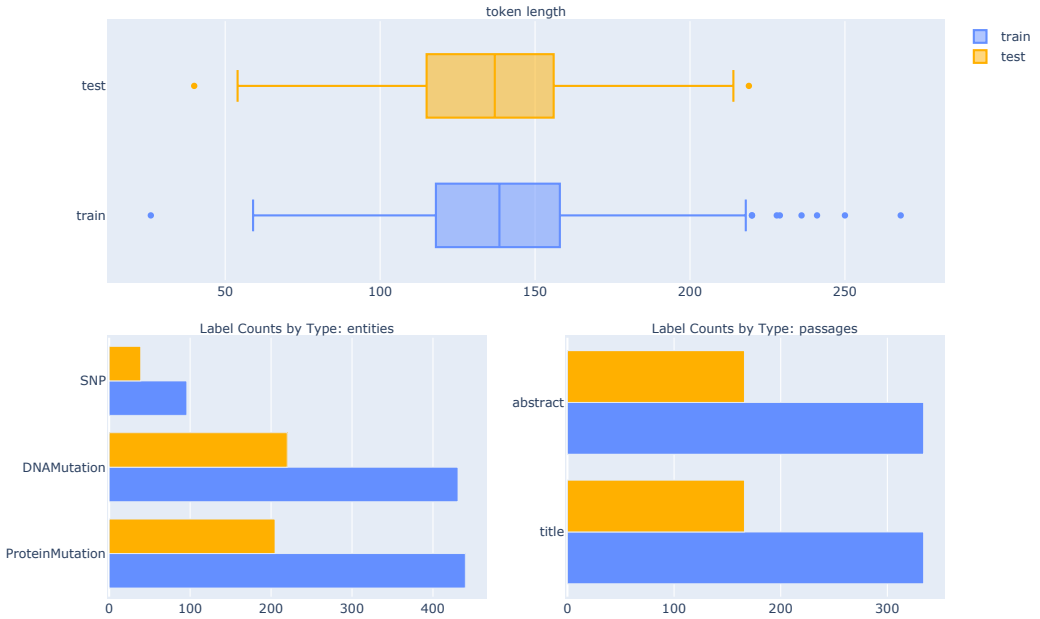

Figure 148: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: This dataset contains 500 PubMed articles manually annotated with mutation mentions of various kinds. It can be used for NER tasks only.

The dataset is split into train(334) and test(166) splits

Homepage: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/`

URL: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition

Schemas: `KB`

Splits: train, test

tmVar v2 Data Card

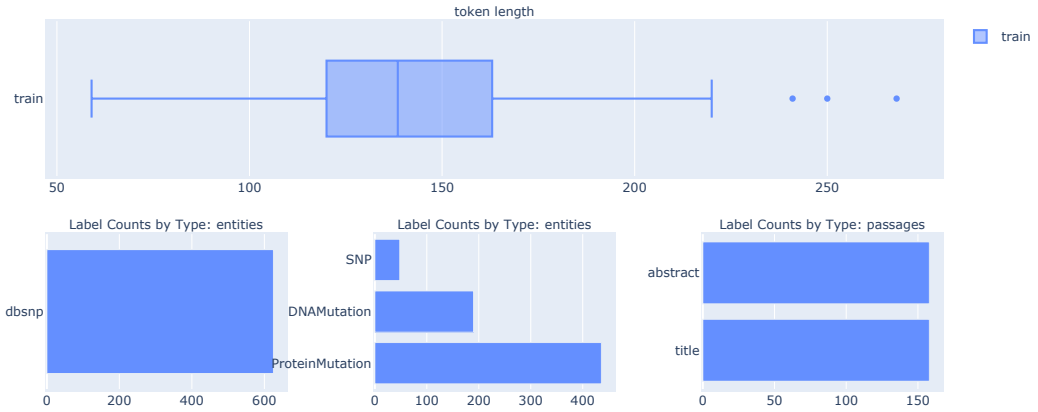

Figure 149: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: This dataset contains 158 PubMed articles manually annotated with mutation mentions of various kinds and dbsnp normalizations for each of them.It can be used for NER tasks and NED tasks, This dataset has a single split

Homepage: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/`

URL: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/`

Licensing: UNKNOWN

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train

tmvar v3 Data Card

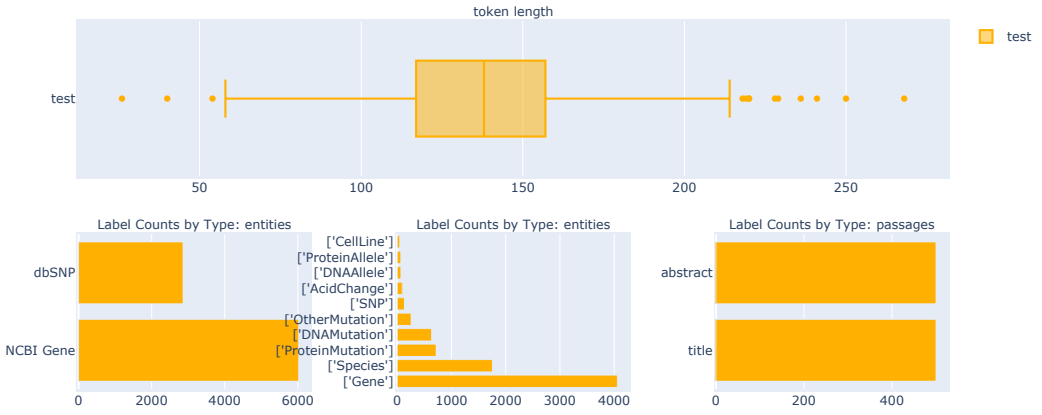

Figure 150: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description This dataset contains 500 PubMed articles manually annotated with mutation mentions of various kinds and dbsnp normalizations for each of them. In addition, it contains variant normalization options such as allele-specific identifiers from the ClinGen Allele Registry. It can be used for NER tasks and NED tasks, This dataset does NOT have splits

Homepage: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/`

URL: `https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/`

Licensing: UNKNOWN

Languages: English

Tasks: named entity disambiguation, named entity recognition

Schemas: `KB`

Splits: test

TwADR-L Data Card

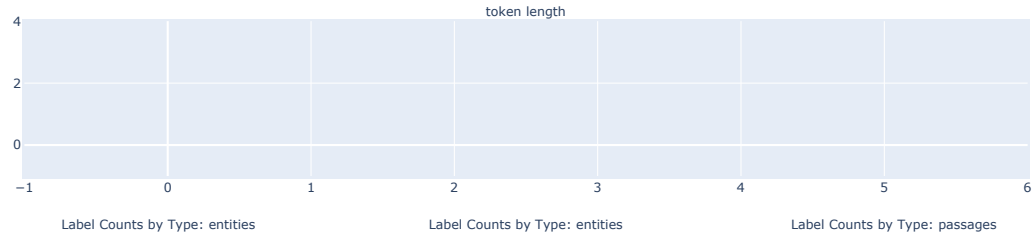

Figure 151: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: The TwADR-L dataset contains medical concepts written on social media (Twitter) mapped to how they are formally written in medical ontologies (SIDER 4).

Homepage: `https://zenodo.org/record/55013`

URL: `https://zenodo.org/record/55013`

Licensing: CC_BY_4p0

Languages: English

Tasks: named entity recognition, named entity disambiguation

Schemas: `KB`

Splits: train, validation, test

UMNSRS (Relatedness) Data Card

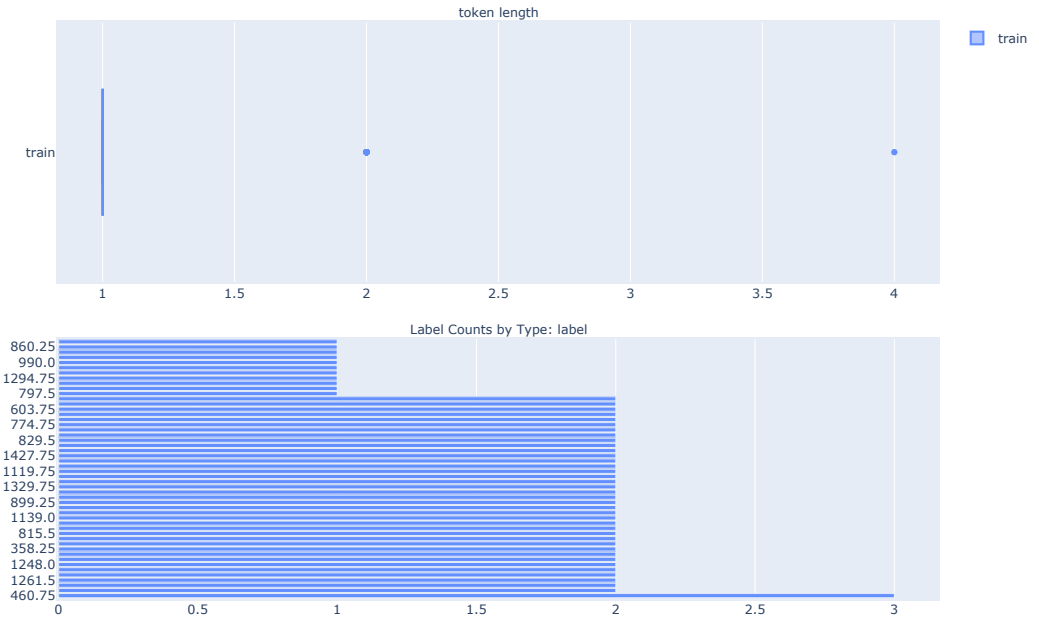

Figure 152: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: UMNSRS, developed by Pakhomov, et al., consists of 725 clinical term pairs whose semantic similarity and relatedness.

The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness.

The following subsets are available:

- similarity: A set of 566 UMLS concept pairs manually rated for semantic similarity (e.g. whale-dolphin) using a continuous response scale.

- relatedness: A set of 588 UMLS concept pairs manually rated for semantic relatedness (e.g. needle-thread) using a continuous response scale.

- similarity_mod: Modification of the UMNSRS-Similarity dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora.

Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 449 pairs.

- relatedness_mod: Modification of the UMNSRS-Relatedness dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora. Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 458 pairs.

Homepage: `https://conservancy.umn.edu/handle/11299/196265/`

URL: `https://conservancy.umn.edu/handle/11299/196265/`

Licensing: CC0_1p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

UMNSRS (Relatedness Mod) Data Card

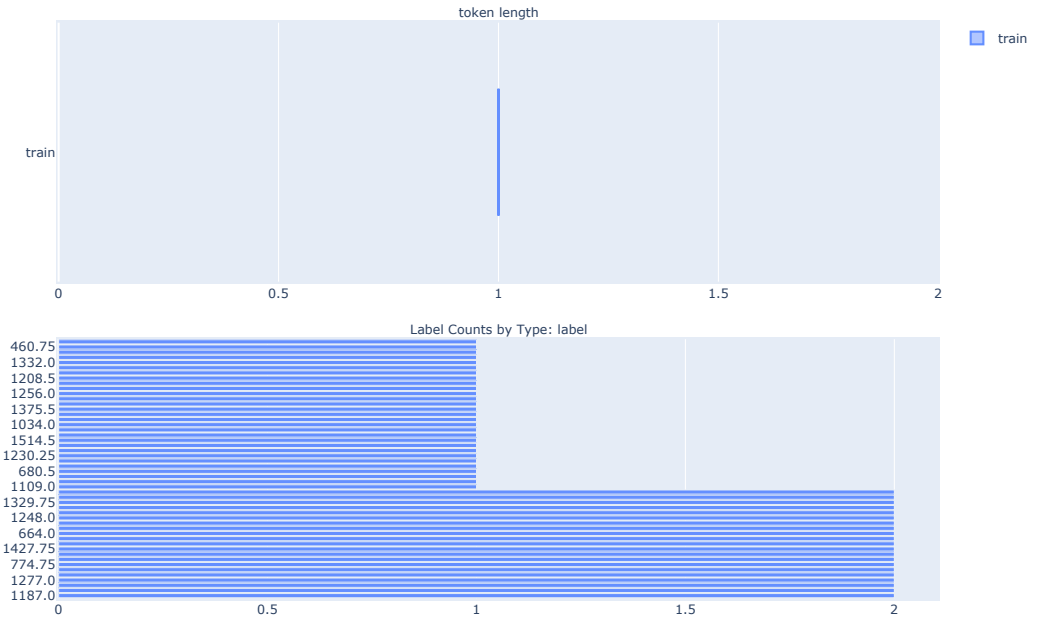

Figure 153: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: UMNSRS, developed by Pakhomov, et al., consists of 725 clinical term pairs whose semantic similarity and relatedness.

The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness.

The following subsets are available:

- similarity: A set of 566 UMLS concept pairs manually rated for semantic similarity (e.g. whale-dolphin) using a continuous response scale.

- relatedness: A set of 588 UMLS concept pairs manually rated for semantic relatedness (e.g. needle-thread) using a continuous response scale.

- similarity_mod: Modification of the UMNSRS-Similarity dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora.

Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 449 pairs.

- relatedness_mod: Modification of the UMNSRS-Relatedness dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora. Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 458 pairs.

Homepage: `https://conservancy.umn.edu/handle/11299/196265/`

URL: `https://conservancy.umn.edu/handle/11299/196265/`

Licensing: CC0_1p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

UMNSRS (Similarity) Data Card

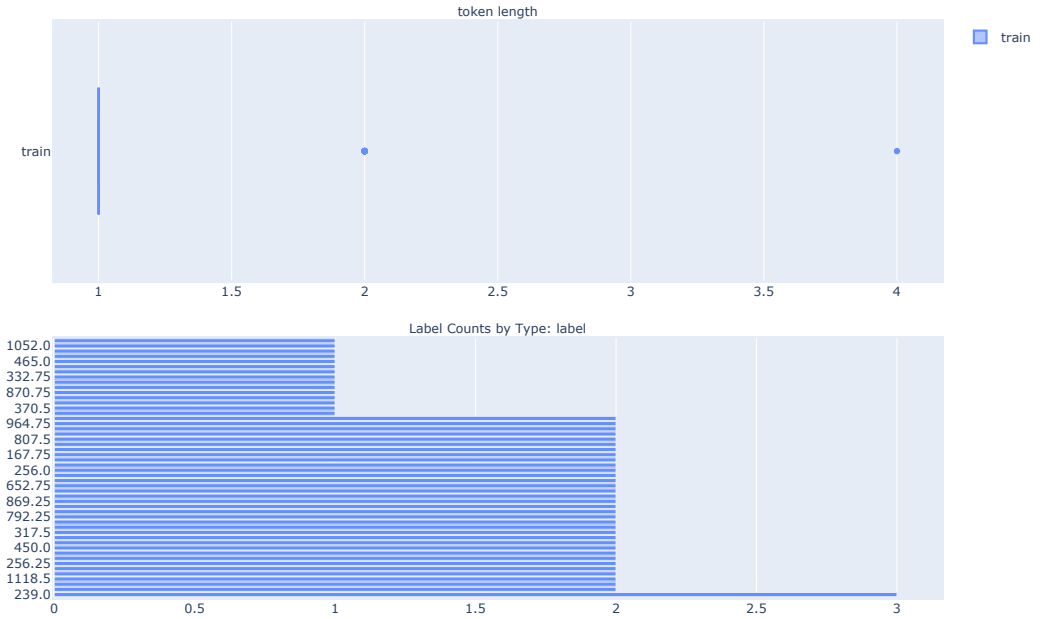

Figure 154: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: UMNSRS, developed by Pakhomov, et al., consists of 725 clinical term pairs whose semantic similarity and relatedness.

The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness.

The following subsets are available:

- similarity: A set of 566 UMLS concept pairs manually rated for semantic similarity (e.g. whale-dolphin) using a continuous response scale.

- relatedness: A set of 588 UMLS concept pairs manually rated for semantic relatedness (e.g. needle-thread) using a continuous response scale.

- similarity_mod: Modification of the UMNSRS-Similarity dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora.

Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 449 pairs.

- relatedness_mod: Modification of the UMNSRS-Relatedness dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora. Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 458 pairs.

Homepage: `https://conservancy.umn.edu/handle/11299/196265/`

URL: `https://conservancy.umn.edu/handle/11299/196265/`

Licensing: CC0_1p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

UMNSRS (Similarity Mod) Data Card

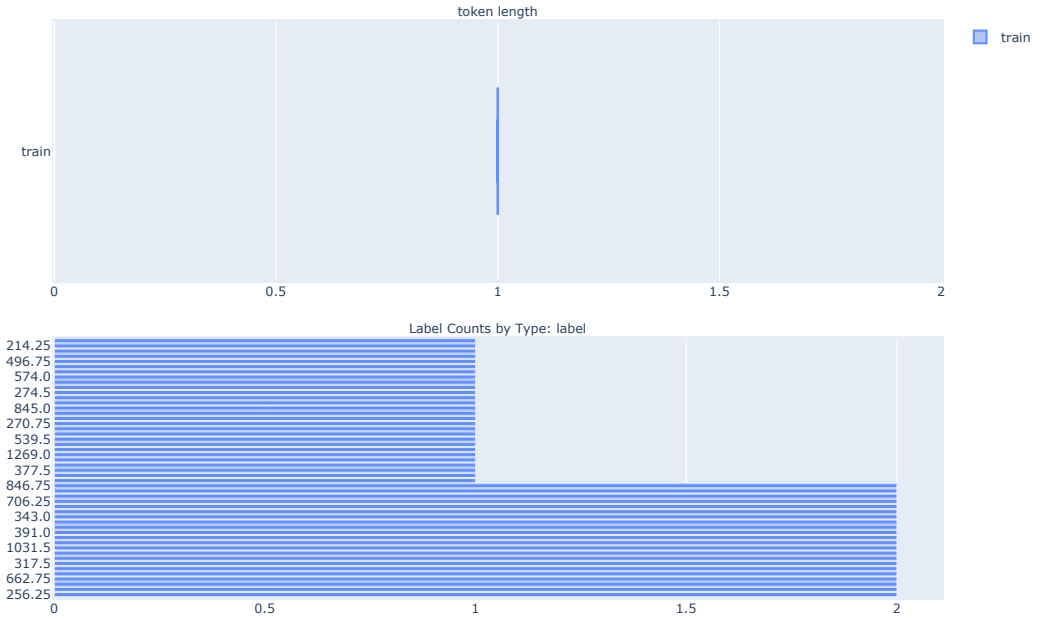

Figure 155: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: UMNSRS, developed by Pakhomov, et al., consists of 725 clinical term pairs whose semantic similarity and relatedness.

The similarity and relatedness of each term pair was annotated based on a continuous scale by having the resident touch a bar on a touch sensitive computer screen to indicate the degree of similarity or relatedness.

The following subsets are available:

- similarity: A set of 566 UMLS concept pairs manually rated for semantic similarity (e.g. whale-dolphin) using a continuous response scale.

- relatedness: A set of 588 UMLS concept pairs manually rated for semantic relatedness (e.g. needle-thread) using a continuous response scale.

- similarity_mod: Modification of the UMNSRS-Similarity dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora.

Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 449 pairs.

- relatedness_mod: Modification of the UMNSRS-Relatedness dataset to exclude control samples and those pairs that did not match text in clinical, biomedical and general English corpora. Exact modifications are detailed in the paper (Corpus Domain Effects on Distributional Semantic Modeling of Medical Terms. Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. Bioinformatics. 2016; 32(23):3635-3644). The resulting dataset contains 458 pairs.

Homepage: `https://conservancy.umn.edu/handle/11299/196265/`

URL: `https://conservancy.umn.edu/handle/11299/196265/`

Licensing: CC0_1p0

Languages: English

Tasks: semantic similarity

Schemas: `PAIRS`

Splits: train

Verspoor 2013 Data Card

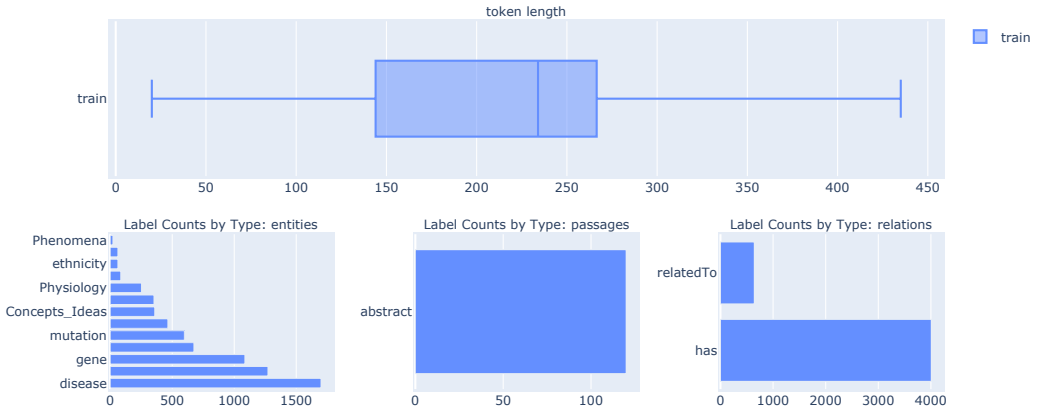

Figure 156: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description: This dataset contains annotations for a small corpus of full text journal publications on the subject of inherited colorectal cancer. It is suitable for Named Entity Recognition and Relation Extraction tasks. It uses the Variome Annotation Schema, a schema that aims to capture the core concepts and relations relevant to cataloguing and interpreting human genetic variation and its relationship to disease, as described in the published literature.The schema was inspired by the needs of the database curators of the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) database, but is intended to have application to genetic variation information in a range of diseases.

Homepage: NA

URL: NA

Licensing: UNKNOWN

Languages: English

Tasks: relation extraction, named entity recognition

Schemas: `KB`

Splits: train