List of existing resources useful for natural language processing in the pharmacovigilance domain

- source: type of source, e.g., scientific paper, abstract, drug leaflet, patient forum, tweet, etc.
- lang = languages: comma-separated list of 2-letter ISO codes
- description: short characterization of the corpus
- noteworthiness: any specific feature of this dataset
- NER: are entities annotated, and for what types of entities
- linking: is entity linking provided, and to what ontologies
- REL: are relations annotated
 - IE = information extraction style: between entity instances (one per pair of entity spans),
 - KB = knowledge-base style: between entities (one per text and pair of [linked] entities),
 - CL = text classification style: presence of a relation between entity types (one per text and pair of entity types); if only one type of relation is considered, this is a binary text classification task
- REL list: if REL is non null, list of annotated relations
- format: CONLL, BRAT, etc.
- size: number of language units such as documents, sentences, words (please no megabytes)
- **publication**: reference to a publication (peer-reviewed rather than preprint)
- URL: URL where the dataset can be downloaded or is described

nar	ne sour	rce lan	g description	noteworthiness	NER	linking	REL	REL list	format	size	publication	URL	
TLC	patiei forum	i ide	dataset annotated with layman expressions: Fachterm, Laienbegriff, Abkürzung		layman terms, including their associated technical terms; technical term with a rather layman term	no	no			BRAT	4000 documents	https://www.aclweb.org/anthology/2020.lrec-1.759/ http://macss	s.dfki.de/data/LREC2020/TLC_v01.tar.gz

From:

https://keepha.lisn.upsaclay.fr/wiki/ - KEEPHA

Permanent link:

https://keepha.lisn.upsaclay.fr/wiki/doku.php?id=resources:existing

Last update: 2021/05/12 17:11

