KG/pre-training resource/corpus

数据集描述&例子&统计信息	使用数据集名称	数据集出处	使用数据集论文	数据集链接	数据集处理代码	备注
Wikipedia abstract数据，abstract里面有人工标注的实体 *The Lord of the Rings* is an *epic high-fantasy* novel written by English author *J. R. R. Tolkien*. Entity Annotations: The Lord of the Rings, Epic (genre), High fantasy, J. R. R. Tolkien 4.4million的Wikipedia articles **Language / Abstracts / Entity links / Triples** Dutch 1,740,494 11,344,612 114,284,973 English 4,415,993 39,650,948 387,953,239 French 1,476,876 11,763,080 116,205,859 German 1,556,343 15,859,142 153,626,686 Italian 907,329 7,705,247 75,698,533 Spanish 1,038,639 11,558,121 111,293,569 All 11,135,674 97,881,150 959,062,859 包括信息：abstract，abstract中的surface form（mention），mention的position，mention对应的KG中的entity（一个链接），mention在语料中出现的次数 {Berlin;http://dbpedia.org/resource/Berlin; 9338} indicates 9.338occurrences of the entity "Berlin" and includes the link tothe corresponding DBpedia resource. 数据集包括两大文件：一个abstract内容，如第一个图；一个surface forms，如这个eg。 **它提供abstract和entity的对齐了没?** **Surface form是什么？DBpedia和Wikipedia的entity id能对上嘛?** **Entity 的size是多少？比abstract多吧？surface form语料是2.6M。**	DBpedia abstract corpus	DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus	Learning Distributed Representations of Texts and Entities fromKnowledge Base, LUKE大概也是用了这个语料，但是预训练的entity vocab size是1M，有点少啊，而且论文对这一块介绍太少了。	http://downloads.dbpedia.org/2015-04/ext/nlp/abstracts/	https://github.com/studio-ousia/ntee	这个工作提到的entity应该是我们常说的mention，也就是数据集论文中的surface form，也就图片中划下划线的mention

english数据。entity和description对齐数据,使用Wikidata dump 2019和Wikipedia数据。description使用的是Wikipedia page中的first section。relation是从Wikidata中抽取的。具体，对于wikidata中的每一个entry，找到他的Wikipedia page，抽取first section作为description。如果找不到Wikipedia page或者page长度小于5个词，则discard。从wikidata中抽取relation时，只保留头尾实体都没有被discard的，并且该relation在wikidata中也有page的。 **wikidata的QID或者entity name和Wikipedia的page是怎么对齐的?**	wikidata5M	KEPLER	KEPLER	https://github.com/thunlp/ERNIE	https://github.com/thunlp/ERNIE
知识图谱KB和free text对齐数据 Wikipedia abstract和wikidata triple对齐数据，指的啥？？ 为抽取T-REx，设计了一套pipeline，document信息使用的DBpedia Abstract Corpus（就是上面提到的那个），triple信息使用的wikidata truthy dump（144M triple） 包括json和NIF两种格式 Json example NIF example	T-REx	T-REx: A Large Scale Alignment of Natural Language withKnowledge Base Triples	K-Adapter 做了一些处理后再使用，产生了T-REx-rc（只包括relation的surface form出现的sentence，丢掉少于50个entity对的relation，最	https://old.datahub.io/dataset/t-rex 或者 https://hadyelsahar.github.io/t-rex/	https://github.com/hadyelsahar/RE-NLG-Dataset

Dataset	#entity	#relation	#training	#validation	#test
FB15K	14,951	1,345	483,142	50,000	59,071
WN18	40,943	18	141,442	5,000	5,000
FB15K-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134
Wikidata5M	4,594,485	822	20,614,279	5,163	5,133

Dataset	Documents / Format	Unique predicates	Aligned Triples	Available
NYT-FB	1.8M sent.	258	39K	partially
TAC KBP	90K sent.	41	122K	closed
Google-RE	60K sent.	5	60K	publicly
FB15K-237	2.7 M patterns	237	2.7M	publicly
Wikireadings	4.7M articles	884	n.a.	publicly

Table 1: Statistics over existing alignments from previous work

Annotator	Documents covered	Alignments	Numerical Alignments	Uniq predicates
NYT-FB	1.8M	39K	None	258
TAC-KBP	0.09M	122K	n.a.	41
*T-REx*_SPO	0.79M	1.2M	21K	336
*T-REx*_NoSub	2.85M	5.2M	561K	642
*T-REx*_AllEnt	**3.09M**	**11.1M**	**350K**	**633**

Table 2: Number of alignments in different datasets

			终得到430个relation和5.5M个sentence）			