A Excluded semantic types

Abbreviation	TUI	Full semantic type name
bird	T012	Bird
dora	T056	Daily or Recreational Activity
edac	T065	Educational Activity
fish	T013	Fish
food	T168	Food
geoa	T083	Geographic Area
gora	T064	Governmental or Regulatory Activity
idcn	T078	Idea or Concept
inpr	T170	Intellectual Product
lang	T171	Language
mnob	T073	Manufactured Object
ocac	T057	Occupational Activity
ocdi	T090	Occupation or Discipline
mcha	T066	Machine Activity
orgt	T092	Organization
phob	T072	Physical Object
phpr	T067	Phenomenon or Process
prog	T097	Professional or Occupational Group
pros	T094	Professional Society
qlco	T080	Qualitative Concept
qnco	T081	Quantitative Concept
rnlw	T089	Regulation or Law
shro	T095	Self-help or Relief Organization
spco	T082	Spatial Concept
tmco	T079	Temporal Concept
vtbt	T010	Vertebrate

Table 1: List of the twenty-six semantic types that were considered non-relevant for biomedical entity linking by the authors. For a full list of all semantic types in the UMLS see: `https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/documentation/SemanticTypesAndGroups.html`

B SPARQL query

```
SELECT ?concept ?conceptLabel ?cui ?article  WHERE {
     ?concept wdt:P2892 ?cui .
    ?article schema:about ?concept .
    ?article schema:isPartOf <https://nl.wikipedia.org/>.

    SERVICE wikibase:label {
       bd:serviceParam wikibase:language "nl"
    }
}
```

Listing 1: SPARQL query for retrieving all Wikidata entities that contain a UMLS CIU and where there exists an article about the entity that is part of the Dutch Wikipedia

C Hyperparameters

Hyperparameter	**Search space**	**Value**
Learning rate (2nd-phase and fine-tuning)		1×10^{-4}
Batch size		512
Weight decay		0.01
Max sequence length		25
Miner margin (λ)		0.2
Random seed		1993
Loss function		MS loss
α in MS loss		2
β in MS loss		50
ϵ in MS loss		0.5
Representation of input string		`[CLS]-token`

Table 2: List of hyperparameters used in the 2nd phase pretraining and fine-tuning steps.

D Hardware details

Hardware	Details
RAM	32GB
GPU	V100 Nvidia GPU
CPU	2vCPU @ 2.2GHz

Table 3: Hardware details that were used for training and evaluating our models on Google Colab Pro.

E Training protocol

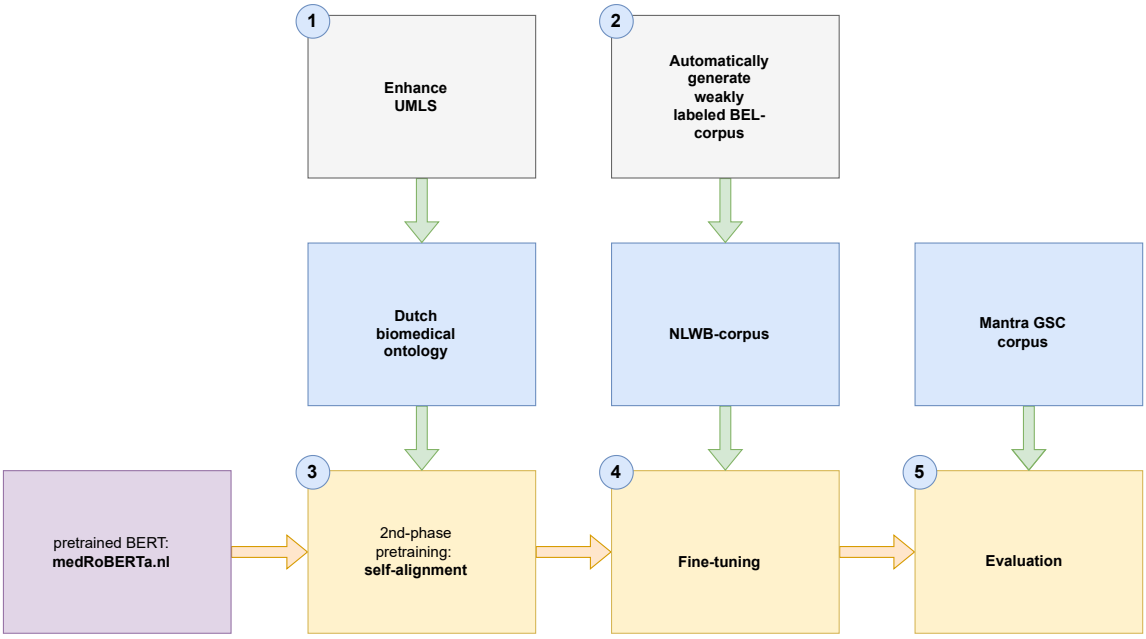

Figure 1: Overview of the training protocol.

F Label quality scores of 100 random samples from NLWB-corpus

id	mention	sentence_id	checked_cuis	quality_score	note
1351	vastendagen	1303	C0015663	1	NaN
1260	toeval	1214	C0036572	1	NaN
1449	afscheiding	1397	C0038984	2	C0038984 codes for 'sweat'
863	antibiotische	833	C0003232	1	(C0282638, Profylaxe, antibiotische) could also
466	PIP	450	C0224637	1	NaN
1083	nekwervels	1043	C3665420	1	NaN
514	kerndelingen	495	C0007610	3	NaN
3758	masochist	3623	C0233869	2	C0233869 is a masochistic personality trait
2129	Progeria	2053	C0033300	1	NaN
7666	Kinine	6899	C0034417	1	NaN
6470	DNA-spoor	5743	C0679560	2	C0679560 codes for genetic test
7422	astmatische	6664	C3714497	2	C3714497 codes for reactive airway disease
1458	traumatisch	1405	C3203533	1	NaN
6550	polioepidemie	5818	C0032371	2	C0032371 codes for Poliomyelitis
6674	Non-hodgkinlymfoom	5939	C0024305	1	NaN
3102	zachte gehemelte	2994	C0700374	2	C0700374 codes for palatum, while there is an
2790	bestraling van kanker	2691	C0261590	1	C1522449 is also correct and maybe better: Ra
625	atherosclerotische	604	C0004153	1	C0003850 is also correct: atherosclerosis
5957	gewrichtspijn	5253	C0003864	2	C0003862 (arthralgia-sign or symptom) would
68	aplastische anemie	66	C0002874,C0178416	1	NaN
2564	Fluoxetine	2470	C0016365	1	NaN
4841	chronische nierziekte	4634	C0022658,C0266292	1	NaN
3477	hoestonderdrukker	3354	C0003449	1	NaN
341	nierschade	330	C0035078,C1565489	1	NaN
3547	ijzersulfaat	3421	C0873006	1	NaN
423	taxol	408	C0144576	1	NaN
719	bevend	695	C0040822	2	C0234369 would have been better (trembling)
2398	anterograde,..., amnesie	2309	C0002622,C0002625	1	NaN
3736	laparoscopisch	3602	C0031150	1	C0393360 is also correct: laparoscopic approach
4727	kalium-ionen	4523	C0032821	1	NaN
214	slaapproblemen	205	C0154564	2	C0851578 would have been better: sleep disord
468	letsels	451	C3263723	1	NaN
3825	Oseltamivir	3689	C0874161	1	NaN
6905	cortex	6165	C0007776	1	NaN
1149	PCP	1107	C0031381	1	NaN
1198	heupbeen	1156	C0019552	1	NaN
6363	syndroom van Gorlin-Goltz	5640	C0004779	1	NaN
2171	neusbijholte	2093	C0030471	1	NaN
6882	hoofdtrauma	6142	C3263723	2	C0744612 would have been more precise: head
318	acute,...,luchtwegen	307	C0009443	1	C4538596 would have also been correct: upper
6161	Bronchitis	5444	C0006277,C0008677	1	NaN
6718	woestijnblindheid	5982	C0155078	3	Not the same thing as far as I know
2285	lyserende werking	2202	C0024348	1	NaN
83	blaas	81	C0005682	1	NaN
5624	TAR-syndroom	4932	C0175703	1	NaN
7098	fibula	6350	C0016068	1	NaN
2747	samentrekken	2649	C0026820	1	NaN
886	knokkelkoorts	855	C0011311	1	NaN
4007	narcose	3849	C0002903	2	C0234447 would have been more precise: narco
1033	gehoorverlies	994	C0018777	2	C1384666 would have been more suited: hearin
252	Leydigcel	243	C0023602	1	NaN
5959	röntgentherapie	5255	C0261590	2	C0043308 would have been more precise: X-ray

7978	trachoma	7203	C0040592,C0153107	1	NaN
943	kloofjes	909	C0221245	1	NaN
7210	TBC-Sanatorium	6459	C0041296	2	C0041296 codes for TBC. C0871286 (sanatorium) or C23502
3304	traumatisch hersenletsel	3189	C0149844	2	C0149844 codes for brain contusion. C0876926 (traumatic br
115	TIA	111	C0917805,C0007787	1	NaN
2778	ovariële	2680	C0029939	1	NaN
6885	DPS	6145	C0683416	1	NaN
2283	borstvlieskanker	2200	C0025500,C0278752	1	NaN
2793	lui oog	2694	C0002418	1	NaN
6068	huidige mens	5353	C0086418	1	NaN
3362	2-butanon	3246	C0066367	1	NaN
150	CFC-11	144	C0077039	1	Dutch ontology does not contani a CFC-11 entry or anything
6081	caustische soda	5366	C0037517	1	NaN
6928	stotterde	6186	C0038131,C0038506	1	NaN
744	anafylactische shock	720	C0002792	1	NaN
555	buikloop	535	C0013369	2	C0011991 (diarrhea) would have been more precise, instead o
2428	nymfomane	2339	C0233619	1	NaN
813	vitamine D2	787	C0014695	1	NaN
7572	breuken	6810	C0016658	1	NaN
4630	MTX	4431	C0025677	1	NaN
414	THC	399	C0039663	1	NaN
3173	seksueel overdraagbare ziekten	3065	C0008149	3	C0008149 codes for Chlamydia
3211	zwaargewonden	3100	C3263723	3	C3263723 codes for (traumatic) injury
6134	Tinnitus	5417	C0040264	1	NaN
3989	waterstofseleniet	3832	C0074280	1	NaN
482	Ibuprofen	465	C0020740	1	NaN
6435	auditieve hallucinaties	5708	C0018524,C0235153	2	C0233762 (hallucinations, auditory) would have been more p
7117	gesclerotiseerd	6368	C0036429	1	NaN
6156	polycystisch ovariumsyndroom	5439	C0032460,C1299574	1	NaN
1120	borderline	1078	C0006012	1	NaN
3517	gliawoekeringen	3393	C0027836	3	C0027836 codes for glial cell
5793	Benzol	5090	C0005036	1	NaN
1603	Wolhynia-koorts	1544	C0040830	1	NaN
2635	botsclerose	2540	C0036429	1	NaN
814	ammoniumpersulfaat	788	C0051723	1	NaN
751	DNA-testen	727	C0679560	1	NaN
2569	rechtersleutelbeen	2475	C0008913	1	NaN
3535	Syfilis	3409	C0007939	2	C0007939 codes for syphilitic chancre. C0039128 would have
4714	vergeten	4510	C0002622,C0002625	2	C0002622 (forgetting) would have been more precise, instead
690	coronamaatregelen	667	C5203670	2	C5203670 codes for COVID19 virus disease
2722	Abortus	2625	C0392535	1	NaN
1585	keel	1526	C0018671	2	C0018671 codes for head and neck neoplasm. C3665375 (thr
3138	winden	3030	C0016204	1	NaN
5056	opgeblazen gevoel	4791	C0013395	2	C0013395 codes for dyspepsia. C0857257 (bloated feeling) w
2856	gram-negatieve bacteriën	2754	C0200966	2	C0200966 codes for the method of gram staining. C0018150
7169	hoogzuilvormig	6419	C0014609	3	C0014609 codes for epithelium
581	stelsels van enkele organen	561	C0460002	1	NaN
4568	Verdikking,.., linkerkamer	4371	C0149721	1	NaN

Label quality score for 100 random samples from the NLWB-corpus graded by the author. Quality score 1 is defined as *correct*, 2 as *similar* and 3 as *incorrect*.