Biomedical Entity Linking for Dutch: fine-tuning a self-alignment BERT model on an automatically generated Wikipedia corpus

Abstract

Biomedical entity linking, a main component in automatic information extraction from biomedical texts, plays a pivotal role in connecting textual entities (such as diseases, drugs and anatomical structures) to their corresponding representations in a structured biomedical knowledge base. The task remains challenging despite recent developments in natural language processing. This report presents the first biomedical entity linking model for the Dutch language. We use medRoBERTa.nl as basemodel and second-phase pretrain through self-alignment on an UMLS-based Dutch biomedical ontology. The model is further fine-tuned on an automatically generated Dutch biomedical entity linking dataset that was derived from Wikipedia. We evaluate our optimal model on the Dutch portion of Mantra GSC-corpus and achieve 55.1% classification accuracy and 71.2% 1-distance accuracy.

1. Introduction

Handling the enormous and ever-expanding volume of biomedical text could benefit from automatic information extraction techniques. Biomedical entity linking (BEL) is the task of linking mentions of biomedical entities in free text to their corresponding canonical form in a knowledge base (Figure 1). This is a necessary step before downstream information extractions tasks can be performed. Applications include automatically categorizing and improving search in medical scientific literature and information extraction from clinical notes and patient fora (Lee et al., 2016). Initial text pattern-based attempts date back to the early 2000s, while modern models incorporate machine-learning algorithms (French & McInnes, 2022). The task remains challenging due to 1) the high diversity in surface form of identical biomedical terms. For example, *MI* and *hartaanval* (heart attack) both belong to the same canonical concept form *myocardinfarct* (myocardial infarction). And 2) the similarity in surface form of different biomedical terms: *candida* and *cardia* refer to a yeast and the heart respectively, while their Levenshtein distance is only two. Another complicating factor arises from the often noisy presentation of free text, including spelling errors and (personal) abbreviations. Moreover, the number of entities in the biomedical domain is very large. The Unified Medical Language System (UMLS), the largest biomedical ontology and composed of various medical vocabularies, contains more than 3.3 million unique concepts (Bodenreider, 2004; Vashishth et al., 2021).

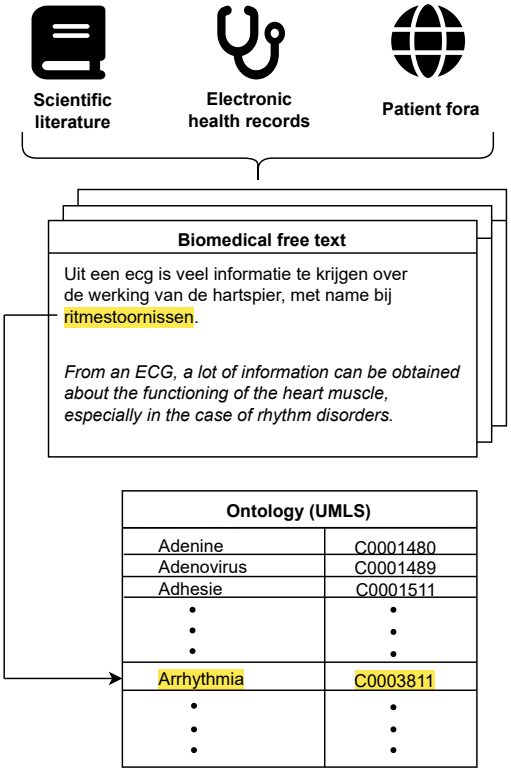

Figure 1. The task of biomedical entity linking. An entity recognition model identifies entities in free text – from e.g. scientific literature, electronic healthcare records or patient fora – that are then passed to the biomedical entity linking model. The biomedical entity linking model associates the new, unseen mention with its corresponding concept from an ontology.

Labeled biomedical entity linking datasets are limited, particularly in languages other than English. In this report, we present NLWB, a weakly labeled Dutch biomedical en-

tity linking dataset that was automatically generated using Wikidata and Wikipedia. By leveraging hyperlinks between Wikipedia articles as entity links, we aimed to obviate the need for expensive, manual annotation by domain experts. We evaluate the quality of the resulting dataset and its effectiveness for training BEL models.

Furthermore, we trained and evaluated the performance of a current state-of-the-art BEL model on Dutch biomedical data. The self-alignment pretraining for BERT (sapBERT) model leverages the occurrence of synonyms in the UMLS for a second phase, self-supervised pretraining step of a previously trained BERT model (Liu et al., 2020). We further fine-tuned this model in a supervised setting on NLWB and evaluated it on the Dutch subset of the Mantra GSC corpus (Kors et al., 2015).

The contributions of this report are as follows: 1) introduction of a method for automatically generating a weakly labeled BEL-dataset in any preferred target language, by combining the UMLS, Wikidata and Wikipedia and thereby obviating the need for manual labelling by a domain expert. 2) introduction of, as fas as we are aware, the first BEL model trained on the Dutch language. 3) Evaluation of the model's performance and generalizability on the Dutch portion of the Mantra GSC dataset.

In the following section, we discuss related work. In Section 3, we formally state the problem of BEL and provide more in depth information about the UMLS and sapBERT. We turn to our methods, mainly the details of our corpus compilation and the 3-stage training protocol of the BERT model, in Section 4. Section 5 contains the results and analysis of our model and our conclusions can be found in Section 6.

2. Related Work

The typical strategy for BEL involves associating a mention with its corresponding concept in a medical ontology, usually the UMLS. Early BEL methods considered this association a dictionary matching problem, that was solved using rule-based algorithms. Commonly, complete pipelines were developed including biomedical named entity recognition (NER), followed by BEL and finally relation extraction (French & McInnes, 2022). Since the mid-2010's, dedicated entity-linking corpora such as ShARe/CLEF and the NCBI dataset were published (Pradhan et al., 2013). This encouraged the development and evaluation of pure BEL-models without possible propagation of errors from the entity-recognition part.

A BEL-model typically involved a candidate generation step followed by candidate ranking (McInnes et al., 2009; D'Souza & Ng, 2015). These rule-based methods are usually fast, but lack semantic and contextual understanding, so they likely fail on linking mentions that are ambiguous. With the development of machine-learning algorithms, the paradigm of how to solve the association step shifted to considering BEL a mapping problem. However, learning the mapping function using traditional supervised learning methods is complicated by the lack of large, labeled datasets for training and the enormous amount of classes (Loureiro & Jorge, 2020).

With representation learning, the need for a labeled dataset can be obviated by leveraging the incorporated knowledge of a medical ontology. A language model such as Bidirectional Encoder Representation from Transformers (BERT) is pretrained on large amounts of text through self-supervised, masked-language modeling (Devlin et al., 2018). Several BERT models pretrained for the biomedical domain exist for English (Lee et al., 2020; Gu et al., 2021). The entity embeddings are then further improved in a second-phase pretraining step by using information from the ontology with strategies like synonym marginalization and self-alignment (Sung et al., 2020; Liu et al., 2020). At inference, a similarity search is performed between the embedding of the new, unseen mention and the precomputed embeddings of all the concepts from the ontology. The mention is then linked to the most similar concept from the ontology.

Improvements have been attempted by incorporating context in the second-phase pretraining step or by using cluster-based inference (Zhang et al., 2021; Ujiie et al., 2021; Angell et al., 2020). More recently, generative language models have also been explored for the task of BEL (Yuan et al., 2022).

3. Background

BEL is the task of mapping mentions in text documents to canonical concepts in a given ontology. A mention is a string that describes some sort of entity in natural language. A concept is a word or phrase that is clearly defined in the ontology and has a unique identifier. Mentions and concepts can either be real world entities or abstract concepts. We formally define the task of BEL as follows:

Problem definition Given a biomedical ontology O consisting of n concepts $O = \{c_1, c_2, ..., c_n\}$, a document D that contains a set M of p biomedical mentions $M = \{m_1, m_2, ..., m_p\}$, the task of BEL is to learn a mapping $M \rightarrow O$ that maps the mention $m_j \in M$ to the corresponding concept $c_i \in O$ that it refers to.

3.1. Unified Medical Language System

The Unified Medical Language System (UMLS) is a large and comprehensive biomedical ontology created and maintained by the US National Library of Medicine. It is a collection of over 160 vocabularies, containing more than

15 million entries in 27 different languages.[1] It maps entries from different databases and terminologies to around 3.3 million unique concepts, that are identified by their Concept Unique Identifier (CUI). The Dutch portion contains around $290,000$ terms. The UMLS also contains data on 54 types of semantic relations between concepts, both hierarchical (e.g. 'is a') and non-hierarchical (e.g. 'is conceptually related to').

3.2. Self-alignment pretraining BERT

The main challenge of BEL in a representation learning setting is the quality of the entity embeddings (Basaldella et al., 2020). Self-supervised learning with masked language modelling on medical data has improved BEL, but does not lead to a well separated representation space (Liu et al., 2020).

Self-alignment pretraining improves the embeddings of a pretrained BERT model, by self-aligning synonymous entries from an biomedical ontology. Formally, the goal of self-alignment is to learn a function $f(.;\theta) : O \rightarrow \mathbb{R}^d$ that is parameterized by θ and where O represents the set of terms in an ontology. In sapBERT, f is modelled by a BERT model with the output `[CLS]` token as embedding representation of the input term c. The similarity of two concepts, $< f(c_i), f(c_j) >$ can be estimated by taking the cosine similarity. During the training procedure, *online hard triplet mining* is used for generating informative pairs that are used for contrastive learning. From each mini-batch, a random anchor term c_a is drawn. Together with a positive match – or synonym – c_p and a negative match c_n, the triplet (c_a, c_p, c_n) is formed. Informative triplets are generated by specifically choosing positive matches with very dissimilar embeddings and negative matches with embeddings that are nearly similar. Formally, triplets are selected that violate the following condition:

$$||f(c_a) - f(c_p)||_2 < ||f(c_a) - f(c_n)||_2 + \lambda \qquad (1)$$

where λ is a pre-set margin. The mining of informative triplets only is useful for improving the embeddings, since otherwise non-informative triplets would dominate the training process due to the enormous size of the ontology (Liu et al., 2020). The Multi-Similarity loss function is used for pulling the embeddings of positive pairs closer and pushing the embeddings of negative pairs further apart (Wang et al., 2019). This process leads to a better separated representation space by leveraging the semantic biases of synonymy relations in the ontology.

[1] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

4. Methods

Due to the need for expensive, manually labelling by domain experts, BEL datasets are not broadly available, especially in languages other than English. We introduce a method for automatically generating a weakly labeled BEL dataset in any given target language, by leveraging the structured knowledge source Wikidata, the UMLS and inter-article hyperlinks on Wikipedia. Since the UMLS is considered somewhat noisy, we first clean and enhance the Dutch subset of the UMLS and generate a Dutch biomedical ontology specifically tailored for BEL tasks.

4.1. Enhancing the UMLS

Roughly 1.7% of the UMLS, comprising $290,056$ terms, is in the Dutch language. However, there is variability in the quality of the records. Entries like:

`voortgangsrapport:bevinding:moment:poliklinisch:document:endocrinologie`
(report:finding:date:polyclinical:document:endocrinology)

are not uncommon. Partly following the Dutch medical concepts[2], we created a cleaned, UMLS-based Dutch biomedical ontology in several steps.

Starting out with the $290,056$ Dutch terms from the 2023 international UMLS release, we first removed vocabularies `LNC-NL-NL` and `ICPC2ICD10DUT`. Both contain too formal, non-informative terms for the task of BEL. Several typos found by the UMCU in terms from the remaining vocabularies were manually corrected. Vocabularies `ICD10DUT` and `MDRDUT` contain terms that include descriptive words as found in definitions, such as `niet-gespecificeerd` (non-specified). We removed those descriptive substrings as they are usually not found in free text. Also, duplicate entries were dropped, irrespective of capitalization. We added the Dutch SNOMED vocabulary, as this is not included in the UMLS. Since the US SNOMED is included in the UMLS, we were able to match Dutch to English terms on their SNOMED ID, and subsequently assign them their corresponding UMLS ID's (CUI's). Several US SNOMED ID's link to more than one UMLS CUI, those ambiguous terms were dropped. Entries linked to one of twenty-six semantic types that were subjectively considered non-relevant for BEL by the authors, such as *Birds* and *Geographic areas* were dropped.[3] Finally, English drug names were added from the `ATC`, `DRUGBANK` and `RXNORM` vocabularies, since they are occasionally used in the Dutch language.

The newly created Dutch biomedical ontology contains $752,536$ terms sourced from 11 vocabularies, all linked

[2] https://github.com/umcu/dutch-medical-concepts

[3] Appendix A: excluded semantic types

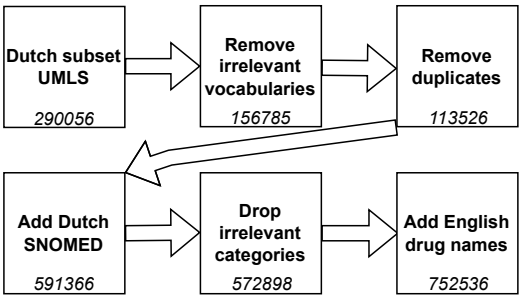

Figure 2. Flow diagram of ontology enhancement. The remaining number of entries are denoted in *italic*.

to one of the $366,071$ distinct concepts. On average, each term is associated with one synonym, but the distribution is heavily right skewed (25% percentile is 0 synonyms per term / 75% percentile is 2 synonyms per term). In Table 1 we show the semantic group distribution of the ontology and the corpora that we will use in the training and evaluation. The semantic groups are not classes in our classification problem, but rather a categorization of the classes. The four largest groups – disorders (DISO), chemicals & drugs (CHEM), procedures (PROC) and anatomy (ANAT) – make up for 97% of the terms in the ontology.

Since the UMLS and SNOMED are licensed, we can not distribute our ontology. However, comprehensive details of all steps are provided in a Python Notebook on Github.[4] The ontology can be reproduced after requesting a UMLS and SNOMED license.

4.2. Corpus compilation

For the automatic generation of our weakly labeled dataset `NLWB`, we combined our enhanced ontology, with textual data from Wikipedia and structured data from Wikidata. Wikidata is a collaboratively edited multilingual knowledge graph that acts as central storage for structured data of its Wikimedia sister projects including Wikipedia (Vrandečić & Krötzsch, 2014). Relevant data can be obtained from Wikidata through SPARQL queries. We retrieved all $4,519$ Wikidata entities that are assigned a UMLS CUI and are linked to a Wikipedia page that is part of the Dutch Wikipedia collection.[5]

We parsed and sentencized all pages from the Dutch Wikipedia dump of March 2023[6] using Python 3.11 and SpaCy sentencizer with the Dutch `nl_core_news_sm` pipeline. We then collected all $51,693$ sentences that contain a hyperlink to one of other $4,519$ Dutch Wikipedia articles, that on their turn are linked to a Wikidata entity with a UMLS CUI. On average, a sentence contains 18 (± 9) tokens. $53,960$ (0.06%) of the tokens in the collection are a biomedical entity mention that is linked to a UMLS CUI (Table 2).

For the NLWB-NO (No-duplicates, Ontology-filtered) subset, we kept each first unique mention but dropped their duplicates. Also, only mentions that link to a CUI that is present in our ontology are included. The NLWB-NO contains $2,400$ unique mentions from $2,307$ sentences. $1,751$ mentions are unseen by our model in the sapBERT training phase as they are not present in our ontology. The mentions map to $1,086$ unique CUI's that are all included in our ontology.

In Table 1, we observe that the mentions in the train- and validation set of NLWB-NO are relatively similarly distributed over the semantic groups as the concepts in the ontology, except for procedures. Procedures are possibly terms more utilized by medical experts only, compared to disorders, chemicals & drugs and anatomical terms, which could explain its lower prevalence on Wikipedia.

The code for parsing the Wikipedia dump and creating the corpus is available on Github.[7] The NLWB corpus and NLWB-NO subset are available in XML format and can be downloaded from Github.[8]

4.3. Basemodel: medRoBERTa.nl

We use a RoBERTa-derived language model that was pretrained from scratch on hospital notes from Dutch electronic healthcare records as basemodel. MedRoBERTa.nl was pretrained on nearly 10 million anonymized hospital notes obtained from the Amsterdam University Medical Centres (Verkijk & Vossen, 2021). The model is distributed with uninitialized head layers, allowing for fine-tuning on specific tasks.

4.4. Second-phase pretraining: self-alignment

Training data for the self alignment pretraining step was generated from the cleaned Dutch medical ontology. We generated a text file with positive pairs in the form of:

`CUI||concept 1||concept 2,`

where concept 1 and concept 2 are synonyms, so associated to the same CUI in the ontology. If more than 2 concepts are associated to the same CUI, all pairwise combinations were traversed and added. The text file was used for sampling the positive pairs during the contrastive learning step

[4]https://github.com/fonshartendorp/dutch_biomedical_entity_linking
[5]The SPARQL query is provided in Appendix B
[6]https://dumps.wikimedia.org
[7]https://github.com/fonshartendorp/dutch_biomedical_entity_linking
[8]https://github.com/fonshartendorp/dutch_biomedical_entity_linking/blob/main/NLWB.xml

Table 1. Semantic group distributions of the ontology, train- and validation set of the no-duplicates, ontology-filtered subset from the NLWB corpus (NLWB-NO) and Mantra GSC. DISO stands for *disorders*, CHEM for *chemicals & drugs*, PROC for *procedures*, ANAT for *anatomy*, LIVG for *living beings*, PHEN for *phenomena*, DEVI for *devices*, PHYS for *physiology* and ACTI for *activities & behaviors*, OBJC for *Objects*, GENE for *genes & molecular sequences*, OCCU for *occupations* and CONC for *concepts & ideas*. 1559 terms in the ontology are not assigned a semantic group (other).

		ONTOLOGY		NLWB-NO TRA.		NLWB-NO VAL.		MANTRA-NO	
GROUP	EXAMPLE	COUNT	PERC.	COUNT	PERC.	COUNT	PERC	COUNT	PERC.
DISO	MS *(multiple sclerosis)*	310057	41.3	957	49.8	224	46.7	149	39.3
CHEM	NEUPRO	185096	24.6	402	20.9	108	22.5	66	17.4
PROC	DIALYSE *(dialysis)*	124345	16.6	90	4.7	20	4.2	68	17.9
ANAT	HEUP *(hip)*	108622	14.5	391	20.4	105	21.9	33	17.4
LIVB	PATIËNT *(patient)*	7586	1.0	14	0.7	6	1.2	29	7.7
PHEN	LICHT *(light)*	5997	0.8	4	0.2	1	0.2	7	1.8
DEVI	IUD'S	3153	0.4	3	0.2	0	0.0	5	1.3
PHYS	GROEI *(growth)*	3125	0.4	33	1.7	11	2.3	19	5.0
ACTI	MACHT *(power)*	1053	0.1	0	0.0	0	0.0	1	0.3
OBJC	STOF *(fabric)*	678	0.1	13	0.7	3	0.6	2	0.5
GENE	CODON	497	0.1	3	0.2	2	0.4	0	0.0
OCCU	GENOMICS	464	0.1	10	0.5	0	0.0	0	0.0
CONC	RETENTIE *(retention)*	304	0.0	0	0.0	0	0.0	0	0.0
OTH.		1559	0.2	0	0.0	0	0.0	0	0.0
TOTAL		**752536**		**1920**		**480**		**379**	

Table 2. Corpora statistics. The NLWB corpus contains many duplicate mentions that occur in different contexts. The NLWB-NO subset contains no duplicate mentions and only links to CUI's that have an entry in our ontology. We created a similar subset of the Mantra GSC corpus (Mantra-NO).

	NLWB	NLWB-NO	MAN.-NO
SENTENCES	51515	2307	166
AVG. #TOK/SENT	18	20	17
MENTIONS	53781	2400	379
UNIQUE MENTIONS	3201	2400	379
UNSEEN MENTIONS	49497	1751	214
CUI'S	56141	2758	402
UNIQUE CUI'S	1334	1086	359
UNLINKABLE CUI'S	47548	0	0

for improving the pretrained BERT embeddings. Negative pairs were sampled online by randomly drawing a concept from the ontology that is not linked to the same CUI. Both the negative and positive pairs must violate the minimum margin condition in Equation 1.

The Multi-Similarity loss is used for re-aligning of the embeddings with its parameters set to the same values as in (Liu et al., 2020).[9] A learning rate of 0.0001 with a weight decay of 0.01 was used for $\{0, 1, 3, 10\}$ epoch(s) with a batch size of 512. The miner margin λ was set to 0.2. The `[CLS]` token was used as representation of the input concept.

[9]Appendix C

The model was built in Pytorch 2.1.0, mostly based on code from (Liu et al., 2020)[10]. Training and evaluation was performed on the Google Colab servers.[11]

4.5. Fine-tuning

Fine-tuning on the NLWB corpus was performed in a similar manner. Now, the positive pairs were generated by combining mentions, linked concepts and their corresponding CUI from the labeled dataset:

```
CUI||mention||linked concept.
```

The hyperparameters were set to the same values as in Section 4.4.[12] We fine-tune for $\{0, 1, 3, 10\}$ epoch(s), building upon the pretrained models from the preceding step.

4.6. Inference

All concepts from the ontology are fed to the trained model, generating a set of precomputed embeddings. At inference, a new, unseen mention is also fed to the trained model and a nearest neighbour search can be performed with the precomputed embeddings. The new mention is assigned the CUI of the most similar embedding from the ontology. Since a nearest neighbour search on $752,536$ items is computationally expensive, a FAISS index was built from the precomputed embeddings instead. FAISS is a library for efficient simi-

[10]https://github.com/cambridgeltl/sapbert

[11]https://colab.research.google.com

[12]Appendix C

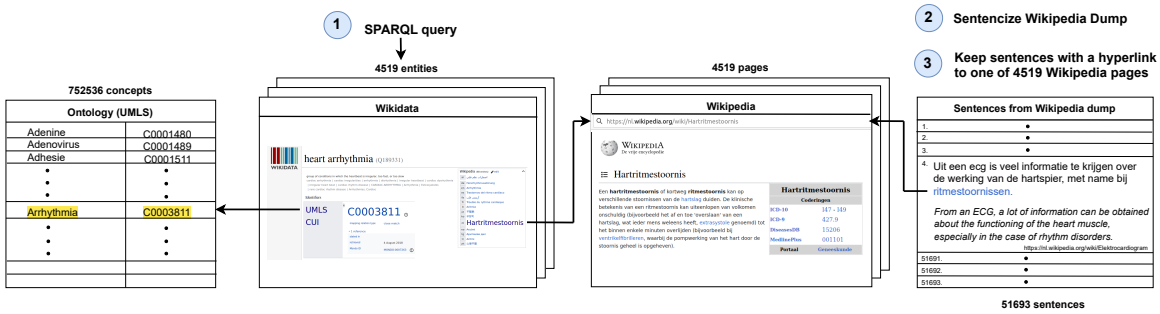

Figure 3. NLWB corpus compilation. All Wikidata entries with a linked Dutch Wikipedia page and a UMLS CUI that is in our ontology are retrieved using SPARQL. Then, all sentences from the Wikipedia March 2023 dump are parsed and selected if they contain a hyperlink to one of the collected Wikipedia pages.

larity search of dense vectors.[13] For memory purposes, the precomputed embeddings are first compressed by using only their first 256 principal components.

4.7. Evaluation data and metrics

We evaluated our method on the Dutch subset of the Mantra GSC corpus. The Mantra GSC corpus is a by domain-experts annotated gold-standard corpus originally created for biomedical concept recognitions in languages other than English (Kors et al., 2015). The texts are sourced from MEDLINE titles and drug labels. The biomedical entities are also annotated with a UMLS CUI, that we use as gold labels for our linking model. Since our ontology does not contain all UMLS CUIs, we use the NLWB-NO and Mantra-NO subsets that contain only mentions that link to a CUI that is included in our ontology. In both corpora, duplicate mentions were also dropped since our model is not context-aware. Table 2 shows the corpora statistics of NLWB-NO and Mantra-NO. The Mantra-NO subset contains 379 mentions from 166 sentences. The sentences are slightly shorter, on average 17 tokens per sentence, and more entities are annotated per sentence.

For finding the optimal number of sapBERT- and fine-tune training epochs, we performed a hyperparameter optimization on the train set of NLWB-NO and validated on its validation set. In the evaluation setting, we fine-tuned our optimal model on the full NLWB-NO subset and evaluated on the Mantra-NO corpus.

We use classification accuracy as the primary metric, in addition to the 1-distance accuracy. For the 1-distance accuracy, predictions are scored correct if they are a 1-distance UMLS relation away from the gold label.

[13]https://github.com/facebookresearch/faiss/wiki

Table 3. Label quality score of the NLWB-corpus. 100 mentions were randomly sampled from the NLWB corpus and manually graded on their label quality by the author. The samples and their grading are given in Appendix E

LABEL QLT. SCORE	CORRECT	RELATED	WRONG
# OF MENTIONS	71	23	6

5. Results and analysis

We assess the quality of our automatically generated NLWB-corpus. Also, the optimal number of training epochs in the sapBERT- and fine-tune pretraining steps are explored. We then turn to an evaluation of the optimal model on Mantra-NO and perform a brief error analysis.

Quality of NLWB-corpus We randomly sampled 100 mentions from the NLWB-corpus and manually evaluated the correctness of their gold label. The grading was performed by the author using the UMLS browser for quick comparison of UMLS entries.[14] Six labels were found to be incorrect. For example, *woestijnblindheid* (desert blindness: disorientation/confusion when navigating in a desert landscape) is labeled as *snow blindness* (C0155078), which is a medical condition caused by overexposure to ultraviolet light. 23 mentions were found to be linked to a concept that seems similar, but not exactly the same. For example, the mention *nierfalen* (kidney failure) on the Dutch Wikipedia page about Francis I of France[15], is linked to Nefritis (Nephritis), which is related but not the same. The remaining 71 mentions seem to be labeled correctly. The 100 samples and their grading can be found in Appendix E. We also observe some mistakes in the sentencizing step, most commonly the title of a certain paragraph ending up prepended to the first

[14]https://github.com/fonshartendorp/dutch_biomedical_entity_linking
[15]https://nl.wikipedia.org/wiki/Frans_I_van_Frankrijk

sentence of that paragraph.

Hyperparameter tuning In Table 4, we present the results of our hyperparameter tuning. The model with 3 sapBERT epochs and 3 fine-tune epochs performed optimal with a classification accuracy of 32.3% and a 1-distance accuracy of 52.5% on the validation set of NLWB-NO.

Table 4. Hyperparameter tuning results on NLWB-NO validation set. Classification- and 1-distance accuracies for the medRoBERTa.nl basemodel (*italic*) and varying epochs of sapBERT and fine-tuning (FT). **Bold** and underline denote the best and second-best results in that column.

TRAINING EPOCHS		NLWB-NO VALID.	
SAPBERT	FT	ACC	1-DIST
0	*0*	*20.8*	*39.8*
0	1	22.9	42.9
0	3	22.5	45.0
0	10	24.6	45.6
1	0	10.6	17.9
1	1	27.7	49.8
1	3	28.8	48.5
1	10	29.4	<u>51.2</u>
3	0	29.2	51.0
3	1	<u>30.2</u>	51.0
3	3	**32.3**	**52.5**
3	10	29.4	51.0
10	0	26.9	49.2
10	1	25.8	47.3
10	3	28.3	50.6
10	10	29.6	51.0

Main results Our optimal model (3S-3FT) achieves a classification accuracy of 55.1% and a 1-distance accuracy of 71.2% on the Mantra-NO corpus (Table 5). That is a 10.5% and 14.5% improvement respectively, compared to the basemodel (BM).

In Table 5, the results grouped by semantic group are separately shown. We do not observe a clear correlation between the size of the semantic groups in the training data (the ontology and NLWB), and their evaluation performance on Mantra-NO. Out of the four main groups in the training data, in two groups – disorders (DISO) and procedures (PROC) – a (around or more than) average improvement is achieved in accuracy over the basemodel with 10.1% and 18.4% respectively. While our model performs less than average on the other two main groups, namely chemicals & drugs (CHEM) and anatomy (ANAT), with an improvement of only 4.5% and 6.1% in classification accuracy. We note that the numbers of mentions per semantic group in Mantra-NO are too small to derive significant conclusions.

Error analysis We make several observations when reviewing mispredictions made by our optimal model. Due to the sometimes noisy and at some points extremely branched structure of the UMLS, seemingly small differences between prediction and gold label are scored incorrect. For example, mention *advies* (advice) is linked to *voorlichting en advies* (counseling-C0010210) by our optimal model. However, in Mantra GSC, its gold label is given as *adviseren* (advice-C0150600). The prediction is called correct by the 1-distance metric, since a RN ('Relation Narrow') exists between the two in the UMLS.

Table 5. Evaluation results on the Mantra-NO corpus for the basemodel (BM) and our optimal model that was trained for 3 self-alignment epochs + 3 fine-tune epochs (3S-3FT). The results are separately shown per semantic group and their respective number of mentions (#) in the corpus. The semantic groups are not classes themselves, but rather a categorization of the classes. DISO stands for *disorders*, PROC for *procedures*, CHEM for *chemicals & drugs*, ANAT for *anatomy*, LIVG for *living beings*, PHYS for *physiology*, PHEN for *phenomena*, DEVI for *devices*, OBJC for *Objects* and ACTI for *activities & behaviors*. The total micro-average is shown for all 379 mentions.

		ACCURACY		1-DIST ACC.	
GROUP	#	BM	3S-3FT	BM	3S-3FT
DISO	149	50.3	60.4	62.4	80.5
PROC	68	26.5	44.9	41.2	58.0
CHEM	66	48.5	53.0	57.6	66.7
ANAT	33	60.6	66.7	66.7	69.7
LIVB	29	34.5	44.8	48.3	58.6
PHYS	19	47.4	63.2	68.4	78.9
PHEN	7	57.1	83.3	71.4	83.3
DEVI	5	20.0	20.0	20.0	40.0
OBJC	2	0.0	0.0	50.0	100.0
ACTI	1	0.0	0.0	0.0	100.0
TOTAL	**379**	**44.6**	**55.1**	**56.7**	**71.2**

Sometimes, a mention is linked to an on surface form-level similar but semantically slightly different concept from the ontology. For example, mention *cannabis* is linked to the plant genus *cannabis* (C0936079), while its gold label in Mantra GSC is the drug *cannabis* (C0678449). There exists a RO ('relation other') between the two. Also, for example mention *pijnlijke rug* (sore back) is linked to *pijnlijke rug* (sore back-C0863105), but labeled as *rugpijn* (back pain-C0004604).

We furthermore observe a high focus on surface form by our models. For example, mention *oren* (ears) is linked to *ren* (running-C0022646) instead of gold label *oor* (ear-C0013443). Moreover, mentions in all capitals, are often linked to a concept in all capitals. Sometimes to a concept that is on surface form and semantical meaning very different. For example, mention *SOMATOTYPE* is linked to *DOPAMINERGIC AGENTS* (C0013036), while the surface form of its gold label is exactly similar to the mention but lower cased: *somatotype* (C0037669). Lower casing all

concepts in the ontology and newly seen mentions could help, but by doing so some information is lost, for example in abbreviations ('pos' is commonly used for 'positive', whereas 'POS' could mean 'Polycystic Ovary Syndrome').

6. Conclusion

To the best of our knowledge, our work is the first to introduce a biomedical entity linking model in the Dutch language. We also present a method for automatically generating a weakly labeled biomedical entity linking dataset in any preferred target language, by combining the data from a biomedical ontology, Wikidata and Wikipedia pages. Using this method, we introduce the first (weakly labeled) Dutch biomedical entity linking corpus: NLWB. As basemodel we use the BERT-based MedRoBERTa.nl language model, that is pretrained for various information extraction tasks on hospital notes from Dutch electronic health records. This model is further refined using self alignment pretraining and fine-tuned on NLWB. We found that 3 epochs of self alignment pretraining and 3 epoch of fine-tuning resulted in optimal performance on the NLWB-NO validation set. On the NLWB-NO validation set, a validation accuracy of 42.0% and a 1-distance accuracy of 52.9% was achieved. That optimal model scored a 55.1% classification accuracy and 71.2% 1-distance relation accuracy on Mantra-NO, an improvement of 10.5% and 14.5% over the basemodel respectively. The lack of professional annotation of the NLWB corpus shows in the low quality of gold labels, which presumably explains the difference in performance on the NLWB-NO validation set and Mantra-NO. We observe that a significant number of mispredictions are actually very close to the gold label, but the noisy and sometimes extremely branched UMLS can give rise to faulty predictions. Also, we note that our model relies heavily on surface form, which is for example observed by the erroneous linking of upper case mentions to upper case concepts that are otherwise very dissimilar. A context-aware model could further improve performance.

References

Angell, R., Monath, N., Mohan, S., Yadav, N., and McCallum, A. Clustering-based inference for biomedical entity linking. *arXiv preprint arXiv:2010.11253*, 2020.

Basaldella, M., Liu, F., Shareghi, E., and Collier, N. Cometa: A corpus for medical entity linking in the social media. *arXiv preprint arXiv:2010.03295*, 2020.

Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

D'Souza, J. and Ng, V. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 297–302, 2015.

French, E. and McInnes, B. T. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics*, pp. 104252, 2022.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Kors, J. A., Clematide, S., Akhondi, S. A., Van Mulligen, E. M., and Rebholz-Schuhmann, D. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956, 2015.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., Lim, S., Choi, D., Kim, S., Tan, A.-C., et al. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10): e0164680, 2016.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*, 2020.

Loureiro, D. and Jorge, A. M. Medlinker: Medical entity linking with neural representations and dictionary matching. In *European Conference on Information Retrieval*, pp. 230–237. Springer, 2020.

McInnes, B. T., Pedersen, T., and Pakhomov, S. V. Umls-interface and umls-similarity: open source software for measuring paths and semantic similarity. In *AMIA annual symposium proceedings*, volume 2009, pp. 431. American Medical Informatics Association, 2009.

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L. M., Vogel, A., Suominen, H., Chapman, W. W., and Savova, G. K. Task 1: Share/clef ehealth evaluation lab 2013. *CLEF (working notes)*, 1179, 2013.

Sung, M., Jeon, H., Lee, J., and Kang, J. Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*, 2020.

Ujiie, S., Iso, H., and Aramaki, E. Biomedical entity linking with contrastive context matching. *arXiv preprint arXiv:2106.07583*, 2021.

Vashishth, S., Newman-Griffis, D., Joshi, R., Dutt, R., and Rosé, C. P. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of biomedical informatics*, 121:103880, 2021.

Verkijk, S. and Vossen, P. Medroberta. nl: a language model for dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, 11:141–159, 2021.

Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57 (10):78–85, 2014.

Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5022–5030, 2019.

Yuan, H., Yuan, Z., and Yu, S. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. *arXiv preprint arXiv:2204.05164*, 2022.

Zhang, S., Cheng, H., Vashishth, S., Wong, C., Xiao, J., Liu, X., Naumann, T., Gao, J., and Poon, H. Knowledge-rich self-supervision for biomedical entity linking. *arXiv preprint arXiv:2112.07887*, 2021.