Multilingual Medical Entity Recognition and Cross-lingual Zero-Shot Linking with Facebook AI Similarity Search

Marcel Schwarz[†], Kathryn Chapman[†] and Bertram Häussler

IGES Institut GmbH, Friedrichstraße 180, 10117 Berlin, Germany

Abstract

We present our submission for the *LivingNER: Named entity recognition, normalization classification of species, pathogens and food* shared task [1] under the team name IGES. Our submission includes predictions for subtasks 1 and 2, LivingNER-Species NER and LivingNER-Species Norm, respectively. We employ a clinically fine-tuned multi-lingual XLM-RoBERTa encoder for both subtasks. We additionally use a Conditional Random Field classifier for Named Entity Recognition, and perform Named Entity Normalization with the Facebook AI Similarity Search (FAISS) library. A key feature of our system is the use of a multilingual encoder which allows zero-shot, cross-lingual medical entity linking. Our system achieves an F1 score of 88.7 for subtask 1, 6.3 points above the mean across all other team submissions, and 87.4 for subtask 2, 4.7 points above the mean.

Keywords

Medical Entity Recognition, Medical Entity Linking, Zero-shot Linking, Cross-lingual entity linking

1. Introduction

Semantic indexing broadly involves linking texts to an external ontology, with the goal of later retrieval based on concepts a text has been linked to via semantic search. This first step can be realized via a variety of approaches, from treating the task as a multi-label classification problem which assigns labels to the entirety of a document, to first extracting specific information from the text which is used to link to the ontology. An example of the latter method is through a pipeline of Named Entity Recognition (NER) followed by Named Entity Normalization (NEN)/Named Entity Linking (NEL). NER aims to isolate subspans in a larger text which represent an entity of interest, often proper names and locations. NEN involves linking those recognized and extracted entities to an ontology or knowledge graph. Major challenges associated with NEN include ambiguity in the extracted entity (e.g. 'President Bush' can refer to two distinct people), monolingual ontologies (which make it difficult to link texts of other languages to the ontology), large ontologies (which require efficient algorithms to search the ontology in reasonable time), and growing ontologies (which regularly contain new concepts).

IberLEF 2022, September 2022, A Coruña, Spain.

[†] These authors contributed equally.

marcel.schwarz@iges.com (M. Schwarz); kathryn.chapman@iges.com (K. Chapman)
0000-0002-5582-2808 (M. Schwarz); 0000-0003-3221-5527 (K. Chapman)

2. Related Work

Biomedical semantic indexing (BSI) involves linking texts to an external biomedical taxonomy or ontology, such as Medical Subject Headings (MeSH) (https://www.ncbi.nlm.nih.gov/mesh/) for clinical publications, International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) (https://icd.who.int/en) for diagnostic and billing purposes, or the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/taxonomy) taxonomy for classifying organisms.

One of the early semantic indexing competitions dealing specifically with biomedical data was the BIOASQ challenge in 2013. Composed of several subtasks, task 1a, "Large-scale online biomedical semantic indexing," involved indexing PubMed abstracts with MeSH codes, as part of larger biomedical question answering (QA) system [2]. The provided data were annotated at the document level, meaning the MeSH codes correspond to the entire abstract. Similarly, the CLEF eHealth 2019 shared task provided non-technical summaries of animal experiments, annotated at the document level with ICD-10 codes, for BSI [3]. As natural language processing (NLP) has migrated from rule-based and classical machine learning algorithms to more powerful deep learning algorithms, interpreting the decisions made by the models has become more challenging. As such, an emerging field in NLP is known as *explainable AI*, where emphasis is placed on finding evidence to support the decisions made by such models. Starting in 2020, more BSI challenges began providing annotations at the entity-level, thereby encouraging participants to adopt a pipeline of named entity recognition (NER) followed by named entity normalization (NEN)/named entity linking (NEL) [4, 5]. This additionally allows for more transparency in the models, as specific spans of text (entities) versus an entire document are linked to an ontology.

BioSyn [6] performs NEN with the synonym marginalization technique, which trains an encoder to produce similar representations for a given medical entity and its corresponding synonyms (e.g. the representation for 'ibuprofen' is similar to that for 'Advil'). BERN2 [7] performs NER using a transformer-based [8] biomedical language model from [9], and NEN by 1) encoding the medical entities with BioSyn and 2) retrieving the most similar candidates with Facebook AI Similarity Search [10]. Similar to BioSyn, [11] performed self-alignment pretraining on a transformer-based language model to encourage projecting all UMLs synonyms for a given medical concept into same area of a shared vector space, such that 'Cerebrovascular accident' and 'stroke' would have similar embeddings.

3. Data

3.1. LivingNER Dataset

The LivingNER dataset consists of 2000 Spanish-language clinical case reports from 20 medical disciplines. The texts are annotated for the tasks of NER, Normalization and Clinical IMPACT and is split into a training set (1000 texts), a validation set (500 texts), and a test set (500 texts). Here, we further describe the annotated datasets for the first two tasks in which we participated. The following statistics do not include the test set, since the labels have not been publicly released at the time of writing.

On average, each text has a length of 33 sentences or 3603 characters. The corpus contains

23203 total annotations (15.5 per text on average), out of which 55.6% have the NER-label "SPECIES", while the remaining are labeled as 'HUMAN'. For the normalization task, the annotations have been manually mapped to codes from the NCBI Taxonomy. All spans labeled as 'HUMAN' are labeled with the code '9606' (*homo sapiens*). The entities labeled as 'SPECIES' are linked to 1041 unique codes. This includes two special types of codes:

- **Complex codes**: If several NCBI taxonomy codes were required to map a single annotated mention, the codes are concatenated with a vertical bar. For instance, *microorganism* is mapped to '2|2759|10239'.
- **More general codes**: If the NCBI taxonomy concept was more general than the annotated mention, the modifier 'H' is added to the NCBI taxonomy code. For instance, *baciloscopia* is mapped to '2|H'.

The complex codes make up 6.6% and the more general codes 8.7% of unique 'SPECIES' codes in the dataset. In total, they account for 7.7% (complex codes) and 6.6% (more general) of all 'SPECIES' annotations. The validation set includes 115 (27.7%) unseen codes, which do not appear in the training data. However, these unseen codes are generally infrequent and, thus, only make for 6.3% of all 'SPECIES' codes in the validation data. Overall, the dataset contains 3792 unique labeled text spans and, on average, each code is linked to 3.6 unique spans. However, more than half of the codes have only one unique text span assigned. On the other side of the spectrum, the code '9606' is linked to 608 different spans, while the code '2|H' has the most unique spans for a 'SPECIES' label with 128. Each code appears 12.4 times on average in the dataset. However, 36.3% of the codes only occur a single time. Figure 3.1 shows the distribution of the most common codes in the corpus and Table 1 displays additional information about the 10 most frequent codes. It is important to note that the complex and more general codes are very top-heavy. The two most common more general codes ('2|H' and '2759|H') account for almost 60% of all annotations of that type, while the most common complex code ('2|2759|10239') makes up >40% of the complex annotations. For further information about the LivingNER corpus, refer to the overview paper for the LivingNER shared task [1].

3.2. NCBI Taxonomy

The NCBI Taxonomy Database is a classification and nomenclature for all of the organisms in the public sequence databases curated by the National Center for Biotechnology Information [12]. It currently contains information for about 10% of the described species of life on the planet.

3.3. UMLS

The Unified Medical Language System (UMLS) (https://www.nlm.nih.gov/research/umls/index.html) is a compendium which provides mappings between various biomedical ontologies across a variety of languages [13]. The Metathesaurus in particular is where the mappings between ontologies such as MeSH and ICD-10 can be found. Additionally, it offers textual descriptions of preferred terms for a given concept and their synonyms in a plethora of languages, including English, German, Spanish, French, etc.

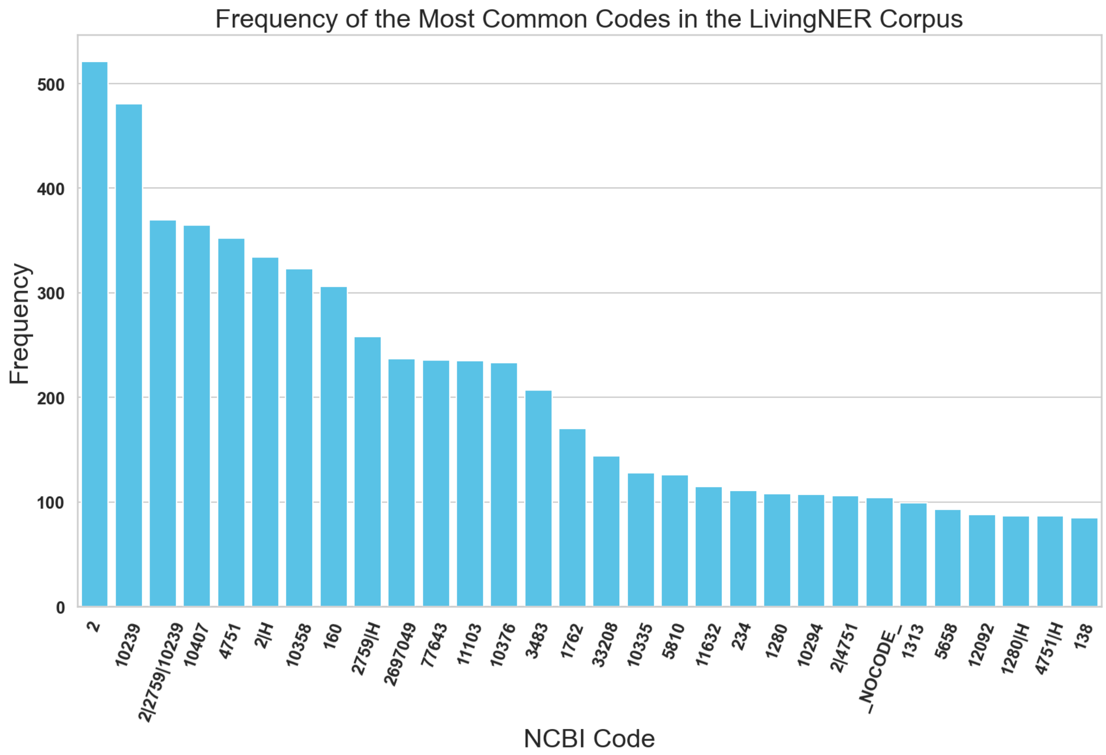

Figure 1: Frequency of the most common codes in the LivingNER corpus. Only the 30 most frequent codes are displayed.

4. Methods

4.1. SAP-BERT

The Self-Alignment Pretraining for Biomedical Entity Representations (SAP-BERT) [11] authors released various models which produce similar medical entity vector representations for synonyms of a given medical concept. They initialized an encoder with the pretrained weights from either PubMed BERT [14] or the multilingual XLM-RoBERTa [15], and performed self-alignment pretraining on all synonyms extracted from UMLS, which contains textual synonyms in a variety of languages. As such, the vectorized representations for 'Hiperglucemia' (English: Hyperglycemia) and 'high blood sugar' would be close together in a shared vector space despite having different surface compositions and being from different languages.

4.2. Named Entity Recognition

Named Entity Recognition is often treated as a sequence-labeling task, where each token in an input text is labeled as either 'B' for 'beginning of entity', 'I' for 'inside entity', or 'O' for 'outside of entity'. In standard NER, these named entities are often proper nouns such as celebrities, cities, or organizations. In the biomedical domain, the task is often referred to as Medical Entity Recognition (MEL) and is commonly restricted to concepts inside an external ontology, such as

Table 1
The ten most common 'SPECIES' codes in the LivingNER corpus. The column 'Text span' shows the span from the corpus which was most often linked to that code. 'English term' shows the preferred term for the NCBI code.

Code	Occurences	Text span	English term
12721	653	VIH	human immunodeficiency virus
2	521	bacteriemia	bacteria
10239	481	viral	viruses
2|2759|10239	370	microorganismos	bacteria|eukaryota|viruses
10407	365	VHB	hepatitis b virus
4751	352	hongos	fungi
2|H	334	baciloscopia	bacteria (more general)
10358	323	CMV	cytomegalovirus
160	306	sífilis	treponema pallidum
2759|H	258	parásitos	eukaryota (more general)

diseases, drugs, or treatments. A deep learning-based technique involves creating vectorized representations of each input token, and training a single linear classifier to output the probabilities that each token is 'B', 'I', or 'O'. These output probabilities can be thought of as 'emission' probabilities, as they are the probability that a given token emits a given label. Transition probabilities, on the other hand, indicate the probability that a given label follows another label. For example, the probability of the 'I' label following the 'O' label is zero, as the 'I' label can only follow the 'B' and 'I' labels. As such, a common challenge with using a non-contextualized linear classifier alone is the output sequence of labels often contains illegal orderings and can require complex post-processing to resolve, even when a contextualized transformer encoder is used to produce the vectorized token representations. [16] found that combining a BERT [17] encoder with a bidirectional long short term memory (LSTM) classifier and a conditional random field (CRF) yielded substantial performance increases for Chinese MER. Therefore, we employ a MER system which utilizes a `SapBERT-UMLS-2020AB-all-lang-from-XLMR` (https://huggingface.co/cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR) encoder with a PyTorch-compatible CRF classifier (https://github.com/kmkurn/pytorch-crf). No layers are frozen during fine-tuning of the encoder/training of the classifier and input documents are split into sentences at both training and inference time for the model inputs.

4.3. Facebook AI Similarity Search

Facebook AI Research provides an open-source library(https://github.com/facebookresearch/faiss) for performing similarity search of dense vectors with GPUs at the billion-scale. The algorithm creates an index, which can contain up to billions of dense vectors, and will retrieve the indices of the top k most similar vectors to a given query vector, using either cosine similarity or L2 (Euclidean) distance, in as little as a few milliseconds.

4.4. Normalization

To perform Medical Entity Normalization (MEN), we first encode all textual terms and their synonyms in the taxonomy of interest using an off-the-shelf `SapBERT-UMLS-2020AB-all-lang-from-XLMR-large` (https://huggingface.co/cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR-large) encoder. For an entity of n length, the output is $n + 2$ contextualized vector representations, as the entity is padded with the `<s>` and `</s>` vectors. The former represents the entire input sequence. For each medical term/synonym, its `<s>` vector is added to a numpy array which is then turned into a searchable index using FAISS.

After extracting a given medical entity via the method described in 4.2, we first re-encode it alone (without its original context) with the same encoder used to create the FAISS index. Then, we retrieve the index of the most similar vector in the FAISS index using Euclidean distance. We define a hyperparameter threshold t, such that if the top match's distance $d \leq t$, the extracted entity is linked to the taxonomy code with which that synonym is associated. If $d > t$, we employ our `ngram_search_loop`. The `ngram_search_loop` iteratively breaks a sequence of whitespace tokens into smaller ngrams in order to link sub-spans within an extracted medical entity to the ontology. The intuition is is the NER step could have extracted either multiple neighboring but distinct medical entities, or neighboring words surrounding a single medical entity, as a single entity which fails to meet the threshold criteria.

4.4.1. `ngram_search_loop`

The `ngram_search_loop` generates all $n - i$grams for the original medical entity, where i starts at 1 and increases until $i == n$. For example, the medical entity 'type 2 diabetes' consists of a single trigram. The `ngram_search_loop` first generates all $n - 1$grams (in this case, bigrams) which are 'type 2' and '2 diabetes'. These are both re-encoded and linking using FAISS is again attempted. The loop continues to generate $n - i$grams, skipping ngrams which contain already linked tokens, until all tokens have been linked to the ontology or $i == n$.

5. Experiments

5.1. Named Entity Recognition

We used the example script `run_ner.py` from the HuggingFace library (https://github.com/huggingface/transformers/blob/main/examples/pytorch/token-classification/run_ner.py). We additionally modified the `RobertaForTokenClassification` class such that the emissions probabilities for the CRF are estimated with a `Linear` layer from PyTorch with `CrossEntropyLoss`, the value of which we denote $Loss_{\text{CE}}$. These emission probabilities are then passed to a CRF token classifier which returns a negative log likelihood (NLL) loss, $Loss_{\text{NLL}}$. The $Loss_{\text{NLL}}$ is scaled by a λ parameter which we set to 0.1, and the final loss is calculated by summing the losses,

$$Loss = Loss_{\text{CE}} + \lambda Loss_{\text{NLL}} \tag{1}$$

We trained the model for 1 epoch on the training data, default parameters otherwise.

5.2. Normalization

The text spans and corresponding labels from the NER layer are used as input to the normalization layer. Spans labeled as ‘HUMAN’ are normalized to the NCBI code ‘9606’ (*homo sapiens*). This represents 43.5% of spans in the training set. The text spans labeled ‘SPECIES’ are re-encoded with SAPBert-XLMR to obtain an embedding without context. For each embedding, the closest synonym is returned from the FAISS index. If the Euclidean distance d is less than or equal to our constant threshold $t = 40$, the code corresponding to the matched synonym is assigned to the text span.

We experimented with three FAISS indexes containing codes and synonyms from different sources:

1. Labeled Spans Index (LSI): This index contains only the codes occuring in the training data. The synonyms for each code are the text spans labeled with that code in the training set. To avoid assigning the same synonym to multiple codes, each synonym was only assigned to the code it was most frequently labeled with.
2. Spanish NCBI Index (SNI): This index is based on the file that was provided for the LivingNER shared task, which contains Spanish translations for the NCBI taxonomy [12].
3. Spanish UMLS Index (SUI): The third index is based on the Unified Medical Language System (UMLS) [13]. It contains the preferred term (English) for all NCBI concepts and the Spanish synonyms from all vocabularies in UMLS.

Only codes from the valid code list provided for the LivingNER shared task [1] were included in the indexes. Statistics about the indexes can be found in Table 2.

Table 2
FAISS Index Information

Index	# Codes	# Synonyms
LSI	887	3006
SNI	1.76M	2.07M
SUI	1.47M	1.99M
SNI+SUI	1.76M	2.46M

6. Results

6.1. Named Entity Recognition

Our SAPBert+CRF model results on the NER subtask are illustrated in Table 3. Our model achieves 6.3 F1 point higher than the mean submission across all participant submitted models.

6.2. Normalization

SNI, SUI, as well as a combination of both indexes, showed very similar performances (Table 4). This could be caused by the relatively high overlap of terms between the two indexes. They

Table 3
Our SAPBert+CRF model performance on the official development and test splits of the LivingNER-Species NER substask 1.

Model	F1	P	R
SAPBert+CRF on Development Set	92.8	94.3	91.4
SAPBert+CRF on Test Set	88.7	91.1	86.4
Average Submission on Test Set	82.4	87.6	80.8

share around 1.6M terms, while 800,000 are unique to one of the indexes. This is not surprising, given that the majority of codes have no synonyms in addition to their preferred term (77% for SNI, 70% for SUI), which is usually an English or a Latin name.

Despite its much smaller size, LSI shows a significantly higher performance than the larger indexes. This is likely due to two reasons. Firstly, LSI contains synonyms for complex codes (e.g. '2|H' or '10407|11103'), while the other indexes can only predict simple codes. These complex codes make up 12.8% of the 'SPECIES' codes in the validation data. Secondly, there is a high overlap between the codes in the training and validation set. In fact, only 6.3% of the annotations are linked to codes which do not occur in the training set. Thus, LSI can identify most text spans, even though it only contains 3006 synonyms. The massive vocabulary of the larger indexes is mostly filled with codes that are irrelevant to the task, which can sometimes act as noise and 'get in the way' of correct codes, by having a similar or even identical synonym. In other words, LSI has a level of task-specific knowledge, i.e. it contains annotations specific to the subject of the corpus and the preferences of the annotators for potentially ambiguous text spans. Often, the predictions of the large indexes are close to the annotated labels but slightly more specific or generic. For example, the span 'enterovirus' is linked to the code '12059' by the large indexes, which is the NCBI code for *enterovirus.* However, the annotated code for the text span is '1193974' (*human enterovirus*). Since the corpus consists of clinical reports about humans, it is often implicit that the human variant of a virus is described and, thus, not explicitly mentioned by the authors. This knowledge is transferred to LSI which contains the corpus-specific synonyms and is therefore able to predict the correct code.

Even the combination of LSI, SNI, and SUI still shows a lower performance than LSI alone (Table 4). The benefits of having a larger vocabulary and containing synonyms for concepts outside of the training set is outweighed by the issue of incorrect synonyms acting as noise. To solve this problem, and combine the strength of task-specific knowledge and an extensive vocabulary, we introduce the approach of index-switching. For this method the smaller index with the most relevant codes (LSI) is searched first. If the closest synonym is below a predefined switch threshold, the corresponding code is predicted as the label. Otherwise, the larger index is searched and the synonym with the lower distance of the two indexes is chosen as the prediction. This approach leads to a slight performance improvement over LSI on the validation data. The increase in recall suggests that the secondary index can add some correct predictions for codes outside of LSI. There was no significant difference in terms of performance between SNI, SUI, and SNI+SUI as a secondary index. Even though the increase in overall performance by adding a second index is relatively low, the system receives a significant boost in its ability to predict a

Table 4
Results on the development set for the Normalization Task. Best Results for each Metric are Indicated in Bold.

Index 1	Index 2	Thr	Sw Thr	Pre	Rec	F1
LSI	-	100	-	**92.5**	86.4	89.3
SNI	-	100	-	72.4	63.3	67.6
SUI	-	100	-	73.8	62.8	67.9
SNI+SUI	-	100	-	71.6	63.4	67.3
LSI+SNI+SUI	-	100	-	85.7	81.6	83.6
LSI	SNI	100	30	92.2	87.7	89.9
LSI	SUI	100	30	92.3	87.7	89.9
LSI	SNI+SUI	100	30	92.3	**87.9**	**90.0**

higher variety of codes. Out of the 559 unique codes in the validation data, LSI correctly predicts 351 codes (62.8%) at least once. LSI+SNI+SUI increases that number significantly to 420 (75.1%), demonstrating its ability at zero-shot entity linking. For a dataset with higher variety, i.e. more annotations with unseen concepts that are not present in the training data, the index-switching approach should lead to an even higher performance gain.

Table 5 illustrates our NEN model's performance on the test set with some of the index configurations from Table 4. Our model achieved an F1 score of 87.4 for subtask 2, 4.7 points above the average.

Table 5
Our NEN model performance on the official test split of the LivingNER-Species NEN substask 2.

Index 1	Index 2	F1	P	R
LSI	-	87.2	90.3	84.4
LSI	SNI+SUI	87.4	89.8	85.1
Avg Submission	-	82.7	84.9	80.7

7. Discussion and Conclusion

Several major challenges of Medical Entity Normalization are addressed by the algorithms we describe in this paper. The utilization of FAISS to link vectorized representations of extracted medical entities to vectorized representations of preferred terms/synonyms of medical entities (of which there are more than there are codes in a given ontology) allows for rapid and efficient linking to large ontologies.

Additionally, the approach can be employed in a zero-shot setting, in that the index can be updated by encoding new, unseen entities without retraining of the system. A potential challenge for this is that the approach relies on an encoder which has been fine-tuned to project synonyms of a same concept near to each other in a shared vector space. For example, 'Advil' (a brand name ibuprofen) can be linked to the concept 'ibuprofen' even if 'Advil' is not in the index as a synonym because the encoder was trained to embed these similarly. As medical science progresses, new drugs and diseases will be added to ontologies and thus retraining of that encoder would likely be necessary eventually. However, the system would still be capable of linking a novel, unseen drug mention extracted from a text to its synonym in an ontology so long as that ontology contains that drug name, e.g. if 'Advil' were a new brand of the new drug 'ibuprofen' and both were encoded and in the searchable index, as RoBERTa models employ subword tokenization to handle unseen words.

Lastly, our model is capable of cross-lingual entity linking, where the entities extracted from a source text can be linked to synonyms in a target ontology of a different language, so long as both of those languages were part of the SAP-BERT-XLMR pretraining and fine-tuning. We would have liked to explore the LivingNER Multilingual datasets, which were machine translated and the token-level annotations automatically transferred to these translated documents. The resulting annotations, we found, were missing several codes present in the original Spanish dataset.

One major challenge in NEN that our model does not address is that of ambiguity, such that we remove the NER-extracted entities from their contexts before re-encoding them and performing linking. Our system would therefore fail to disambiguate an entity such as 'Paris' and properly link it to either Paris Hilton, or Paris, France, where context would likely be vital. Future work could include performing further fine-tuning of the SAP-BERT-XLMR model to perform the linking with contextualized medical entity vector representations.

Acknowledgments

This work was partially funded by the Investitionsbank Berlin (IBB) through the project ProFit.

References

[1] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, D. Estrada, L. Gascó, M. Krallinger, Mention detection, normalization classification of species, pathogens, humans and food in clinical documents: Overview of LivingNER shared task and resources, *Procesamiento del Lenguaje Natural* (2022).

[2] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. Alvers, D. Weißenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A.-C. Ngonga Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, *BMC Bioinformatics* 16 (2015) 138. doi:10.1186/s12859-015-0564-6.

[3] L. Kelly, H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, et al., Overview of the CLEF eHealth evaluation lab 2019, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2019, pp. 322–339.

[4] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in

spanish, corpus, guidelines, methods and results, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*, 2020.

[5] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020, in: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.

[6] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical Entity Representations with Synonym Marginalization, 2020. URL: https://arxiv.org/abs/2005.00239. doi:10.48550/ARXIV.2005.00239.

[7] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, J. Kang, BERN2: an advanced neural biomedical named entity recognition and normalization tool, 2022. URL: https://arxiv.org/abs/2201.02080. doi:10.48550/ARXIV.2201.02080.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2017. URL: https://arxiv.org/abs/1706.03762. doi:10.48550/ARXIV.1706.03762.

[9] P. Lewis, M. Ott, J. Du, V. Stoyanov, Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art, in: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, Association for Computational Linguistics, Online, 2020, pp. 146–157. URL: https://aclanthology.org/2020.clinicalnlp-1.17. doi:10.18653/v1/2020.clinicalnlp-1.17.

[10] D. Danopoulos, C. Kachris, D. Soudris, Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud, in: *SAMOS*, 2019.

[11] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-Alignment Pretraining for Biomedical Entity Representations, 2020. URL: https://arxiv.org/abs/2010.11784. doi:10.48550/ARXIV.2010.11784.

[12] S. Federhen, The NCBI taxonomy database, *Nucleic acids research* 40 (2012) D136–D143.

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

[14] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, 2020. arXiv:arXiv:2007.15779.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, *arXiv preprint arXiv:1911.02116* (2019).

[16] X. Li, H. Zhang, X.-H. Zhou, Chinese clinical named entity recognition with variant neural structures based on BERT methods, Journal of Biomedical Informatics 107 (2020) 103422. URL: https://www.sciencedirect.com/science/article/pii/S1532046420300502. doi:https://doi.org/10.1016/j.jbi.2020.103422.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.